

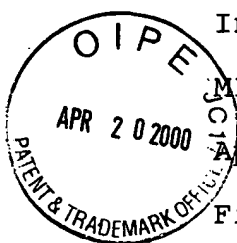
0250

#4

1263.1240

PATENT APPLICATION

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE



In re Application of:)
MICHAEL JAMES TAYLOR ET AL.) : Examiner: Not Yet Assigned
Application No.: 09/532,533) : Group Art Unit: NYA
Filed: March 22, 2000)
For: PROCESSING APPARATUS)
FOR DETERMINING WHICH)
PERSON IN A GROUP IS)
SPEAKING) : April 19, 2000

Assistant Commissioner for Patents
Washington, D.C. 20231

CLAIM TO PRIORITY

Sir:

Applicants hereby claim priority under the
International Convention and all rights to which they are
entitled under 35 U.S.C. § 119 based upon the following
British Priority Applications:

9907103.7, filed March 26, 1999; and

9908546.6, filed April 14, 1999.

Certified copies of the priority documents are
enclosed.

THIS PAGE BLANK (USPTO)

Applicants' undersigned attorney may be reached in our New York office by telephone at (212) 218-2100. All correspondence should continue to be directed to our address given below.

Respectfully submitted,



Attorney for Applicants

Registration No. 28,296
29,296

FITZPATRICK, CELLA, HARPER & SCINTO
30 Rockefeller Plaza
New York, New York 10112-3801
Facsimile: (212) 218-2200

NY_MAIN 76676 v 1

THIS PAGE BLANK (USPTO)



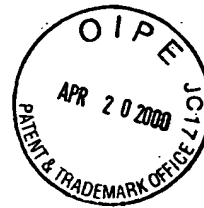
Best Available Copy

The
Patent
Office

091532 533



INVESTOR IN PEOPLE



The Patent Office
Concept House
Cardiff Road
Newport
South Wales
NP10 8QQ


**CERTIFIED COPY OF
PRIORITY DOCUMENT**

I, the undersigned, being an officer duly authorised in accordance with Section 74(1) and (4) of the Deregulation & Contracting Out Act 1994, to sign and issue certificates on behalf of the Comptroller-General, hereby certify that annexed hereto is a true copy of the documents as originally filed in connection with the patent application identified therein.

In accordance with the Patents (Companies Re-registration) Rules 1982, if a company named in this certificate and any accompanying documents has re-registered under the Companies Act 1980 with the same name as that with which it was registered immediately before re-registration save for the substitution as, or inclusion as, the last part of the name of the words "public limited company" or their equivalents in Welsh, references to the name of the company in this certificate and any accompanying documents shall be treated as references to the name with which it is so re-registered.

In accordance with the rules, the words "public limited company" may be replaced by p.l.c., plc, P.L.C. or PLC.

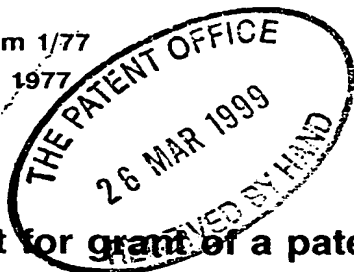
Re-registration under the Companies Act does not constitute a new legal entity but merely subjects the company to certain additional company law rules.

Signed 

Dated 7 April 2000

THIS PAGE BLANK (USPTO)

Patents Form 1/77
Patents Act 1977
(Rule 16)



**The
Patent
Office**

29MAR99 E436117-20 002917
P01/7700 0.00 - 9907103.7

Request for grant of a patent

The Patent Office
Cardiff Road
Newport
Gwent NP9 1RH

| | | |
|----|---|--|
| 1. | Your reference 2647101 | |
| 2. | Patent Application Number | 9907103.7 |
| 3. | Full name, address and postcode of the or of each applicant (<i>underline all surnames</i>) Canon Kabushiki Kaisha 30-2 3-Chome Shimomaruko Ohta-Ku Tokyo Japan Patents ADP number (<i>if known</i>) 0363010003 If the applicant is a corporate body, give the country/state of its incorporation Country: JAPAN State: | |
| 4. | Title of the invention IMAGE PROCESSING APPARATUS | |
| 5. | Name of agent "Address for Service" in the United Kingdom to which all correspondence should be sent Patents ADP number 01826001 | Beresford & Co 2/5 Warwick Court High Holborn London WC1R 5DJ |
| 6. | Priority details Country Priority application number Date of filing | |

Patents Form 1/77

7. If this application is divided or otherwise derived from an earlier UK application give details

Number of earlier of application

Date of filing

8. Is a statement of inventorship and or right to grant of a patent required in support of this request?

YES

9. Enter the number of sheets for any of the following items you are filing with this form.

Continuation sheets of this form 0

Description

48

Claim(s)

10

Abstract

1

Drawing(s)

18

10. If you are also filing any of the following, state how many against each item.

Priority documents

Translations of priority documents

Statement of inventorship and
right to grant of a patent (*Patents form 7/77*)

1 + 2 copies

Request for preliminary examination
and search (*Patents Form 9/77*)

Request for Substantive Examination
(*Patents Form 10/77*)

Any other documents
(*please specify*)

11. I/We request the grant of a patent on the basis of this application

Signature


BERESFORD & Co

Date

26 March 1999

12. Name and daytime telephone number of
person to contact in the United Kingdom

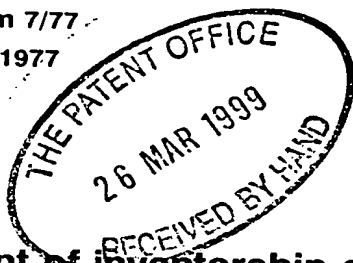
David SPROSTON

Tel:0171-831-2290

Patents Form 7/77

Patents Act 1977

(Rule 15)



**The
Patent
Office**

**Statement of inventorship and of
right to grant of a patent**

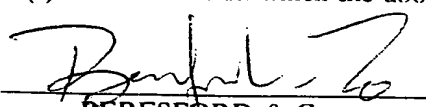
The Patent Office

Cardiff Road

Newport

Gwent NP9 1RH

1. Your reference
2647101
2. Patent Application Number
accompanying application reference 2647101
9907103.7
3. Full name of the or each applicant
Canon Kabushiki Kaisha
4. Title of the invention
IMAGE PROCESSING APPARATUS
5. State how the applicant(s) derived the right from the inventor(s) to be granted a patent
By virtue of the employment of the inventors by Canon Research Centre Europe Ltd, and by virtue of an agreement between Canon Research Centre Europe Ltd and Canon Kabushiki Kaisha dated 1 January 1994.
6. How many, if any additional Patents Forms
7/77 are attached to this form?
11. I/We believe that the person(s) named over the page (and on any extra copies of this form) is/are the inventor(s) of the invention which the above patent application relates to.

Signature  Date **26 March 1999**
BERESFORD & Co
12. Name and daytime telephone number of person to contact in the United Kingdom
David SPROSTON
Tel: 0171-831-2290

Patents Form 7/77

ROWE; Simon Michael
c/o CANON RESEARCH CENTRE
EUROPE LTD
1 Occam Court, Occam Road
Surrey Research Park
Guildford
Surrey GU2 5YJ 076 29595001

TAYLOR; Michael James
c/o CANON RESEARCH CENTRE
EUROPE LTD
1 Occam Court, Occam Road
Surrey Research Park
Guildford
Surrey GU2 5YJ 07587553001

IMAGE PROCESSING APPARATUS

The present invention relates to the processing of image data and sound data to generate data to assist in
5 archiving the image and sound data.

Many databases exist for the storage of data. However, the existing databases suffer from the problem that the ways in which the database can be interrogated to
10 retrieve information therefrom are limited.

The present invention has been made with this problem in mind.

15 According to the present invention, there is provided an apparatus or method in which image data is processed to determine which person in the images is speaking by determining which person has the attention of the other people in the image, and sound data is processed to
20 generate text data corresponding to the words spoken by the person using processing parameters selected in dependence upon the speaking participant identified by processing the image data.

25 The present invention also provides an apparatus or method in which image data is processed to determine at

whom each person in the images is looking and to determine which of the people is speaking based thereon, and sound data is processed to perform speech recognition for the speaking participant.

5

In this way, the speaking participant can be readily identified to enable the sound data to be processed.

The present invention further provides an apparatus or
10 method for processing image data in such a system.

The present invention further provides instructions, including in signal and recorded form, for configuring a programmable processing apparatus to become arranged
15 as an apparatus, or to become operable to perform a method, in such a system.

Embodiments of the invention will now be described, by way of example only, with reference to the accompanying
20 drawings, in which:

Figure 1 illustrates the recording of sound and video data from a meeting between a plurality of participants;

25 Figure 2 is a block diagram showing an example of notional functional components within a processing

apparatus in an embodiment;

Figure 3 shows the processing operations performed by processing apparatus 24 in Figure 2 prior to the meeting
5 shown in Figure 1 between the participants starting;

Figure 4 schematically illustrates the data stored in meeting archive database 60 at step S4 in Figure 3;

10 Figure 5 shows the processing operations performed at step S34 in Figure 3;

Figure 6 shows the processing operations performed by processing apparatus 24 in Figure 2 while the meeting
15 between the participants is taking place;

Figure 7 shows the processing operations performed at step S72 in Figure 6;

20 Figure 8 shows the processing operations performed at step S80 in Figure 7;

Figure 9 illustrates the viewing ray for a participant used in the processing performed at step S114 in
25 Figure 8;

Figure 10 illustrates the angles calculated in the processing performed at step S114 in Figure 8;

Figure 11 shows the processing operations performed at
5 step S84 in Figure 7;

Figure 12 schematically illustrates the storage of information in the meeting archive database 60;

10 Figures 13A and 13B show examples of viewing histograms defined by data stored in the meeting archive database 60;

Figure 14 shows the processing operations performed by
15 processing apparatus 24 to retrieve information from the meeting archive database 60;

Figure 15A shows the information displayed to a user at
step S200 in Figure 14;

20

Figure 15B shows an example of information displayed to a user at step S204 in Figure 14; and

Figure 16 schematically illustrates an embodiment in
25 which a single database stores information from a plurality of meetings and is interrogated from one or

more remote apparatus.

Referring to Figure 1, a video camera 2 and one or more microphones 4 are used to record image data and sound data respectively from a meeting taking place between a group of people 6, 8, 10, 12.

The image data from the video camera 2 and the sound data from the microphones 4 is input via cables (not shown) to a computer 20 which processes the received data and stores data in a database to create an archive record of the meeting from which information can be subsequently retrieved.

Computer 20 comprises a conventional personal computer having a processing apparatus 24 containing, in a conventional manner, one or more processors, memory, sound card etc., together with a display device 26 and user input devices, which, in this embodiment, comprise a keyboard 28 and a mouse 30.

The components of computer 20 and the input and output of data therefrom are schematically shown in Figure 2.

Referring to Figure 2, the processing apparatus 24 is programmed to operate in accordance with programming

instructions input, for example, as data stored on a data storage medium, such as disk 32, and/or as a signal 34 input to the processing apparatus 24, for example from a remote database, by transmission over a communication
5 network (not shown) such as the Internet or by transmission through the atmosphere, and/or entered by a user via a user input device such as keyboard 28 or other input device.

10 When programmed by the programming instructions, processing apparatus 24 effectively becomes configured into a number of functional units for performing processing operations. Examples of such functional units and their interconnections are shown in Figure 2. The
15 illustrated units and interconnections in Figure 2 are, however, notional and are shown for illustration purposes only, to assist understanding; they do not necessarily represent the exact units and connections into which the processor, memory etc of the processing apparatus 24
20 become configured.

Referring to the functional units shown in Figure 2, a central controller 36 processes inputs from the user input devices 28, 30 and receives data input to the
25 processing apparatus 24 by a user as data stored on a storage device, such as disk 38, or as a signal 40

transmitted to the processing apparatus 24. The central controller 36 also provides controller and processing for a number of the other functional units. Memory 42 is provided for use by central controller 36 and other
5 functional units.

Head tracker 50 processes the image data received from video camera 2 to track the position and orientation in three dimensions of the head of each of the participants
10 6, 8, 10, 12 in the meeting. In this embodiment, to perform this tracking, head tracker 50 uses data defining a three-dimensional computer model of the head of each of the participants and data defining features thereof which is stored in head model store 52, as will be
15 described below.

Voice recognition processor 54 processes sound data received from microphones 4. Voice recognition processor 40 operates in accordance with a conventional voice
20 recognition program, such as "Dragon Dictate" or IBM "ViaVoice", to generate text data corresponding to the words spoken by the participants 6, 8, 10, 12. To perform the voice recognition processing, voice recognition processor 54 uses data defining the speech
25 recognition parameters for each participant 6, 8, 10, 12, which is stored in speech recognition parameter store 56.

More particularly, the data stored in speech recognition parameter store 56 comprises data defining the voice profile of each participant which is generated by training the voice recognition processor in a conventional manner. For example, the data comprises the data stored in the "user files" of Dragon Dictate after training.

Archive processor 58 generates data for storage in meeting archive database 60 using data received from head tracker 50 and voice recognition processor 54. More particularly, as will be described below, the video data from camera 2 and sound data from microphones 4 is stored in meeting archive database 60 together with text data from voice recognition processor 54 and data defining at whom each participant in the meeting was looking at a given time.

Text searcher 62, in conjunction with central controller 36, is used to search the meeting archive database 60 to find and replay the sound and video data for one or more parts of the meeting which meet search criteria specified by a user, as will be described in further detail below.

Display processor 64 under control of central controller 36 displays information to a user via display device 26

and also replays sound and video data stored in meeting archive database 60.

Output processor 66 outputs part or all of the data from
5 archive database 60, for example on a storage device such as disk 68 or as a signal 70.

Before beginning the meeting, it is necessary to initialise computer 20 by entering data which is
10 necessary to enable processing apparatus 24 to perform the required processing operations.

Figure 3 shows the processing operations performed by processing apparatus 24 during this initialisation.

15 Referring to Figure 3, at step S2, central controller 36 causes display processor 64 to display a message on display device 26 requesting the user to input the names of each person who will participate in the meeting.

20 At step S4, upon receipt of data defining the names, for example input by the user using keyboard 28, central controller 36 allocates a unique participant number to each participant, and stores data, for example table 80
25 shown in Figure 4, defining the relationship between the participant numbers and the participants' names in the

meeting archive database 60.

At step S6, central controller 36 searches the head model store 52 to determine whether data defining a head model
5 is already stored for each participant in the meeting.

If it is determined at step S6 that a head model is not already stored for one or more of the participants, then, at step S8, central controller 36 causes display
10 processor 64 to display a message on display device 26 requesting the user to input data defining a head model of each participant for whom a model is not already stored.

15 In response, the user enters data, for example on a storage medium such as disk 38 or by downloading the data as a signal 40 from a connected processing apparatus, defining the required head models. Such head models may be generated in a conventional manner, for example as
20 described in "An Analysis/Synthesis Cooperation for Head Tracking and Video Face Cloning" by Valente et al in Proceedings ECCV '98 Workshop on Perception of Human Action, University of Freiberg, Germany, June 6 1998.

25 At step S10, central controller 36 stores the data input by the user in head model store 52.

At step S12, central controller 36 and display processor 64 render each three-dimensional computer head model input by the user to display the model to the user on display device 26, together with a message requesting the user to identify at least seven features in each model.

In response, the user designates using mouse 30 points in each model which correspond to prominent features on the front, sides and, if possible, the back, of the participant's head, such as the corners of eyes, nostrils, mouth, ears or features on glasses worn by the participant, etc.

At step S14, data defining the features identified by the user is stored by central controller 36 in head model store 52.

On the other hand, if it is determined at step S6 that a head model is already stored in head model store 52 for each participant, then steps S8 to S14 are omitted.

At step S16, central controller 36 searches speech recognition parameter store 56 to determine whether speech recognition parameters are already stored for each participant.

If it is determined at step S16 that speech recognition parameters are not available for all of the participants, then, at step S18, central controller 36 causes display processor 64 to display a message on display device 26
5 requesting the user to input the speech recognition parameters for each participant for whom the parameters are not already stored.

In response, the user enters data, for example on a
10 storage medium such as disk 38 or as a signal 40 from a remote processing apparatus, defining the necessary speech recognition parameters. As noted above, these parameters define a profile of the user's speech and are generated by training a voice recognition processor in
15 a conventional manner. Thus for example, in the case of a voice recognition processor comprising Dragon Dictate, the speech recognition parameters input by the user correspond to the parameters stored in the "user files" of Dragon Dictate.

20

At step S20, the data input by the user is stored by central controller 36 in the speech recognition parameter store 56.

25 On the other hand, if it is determined at step S16 that the speech recognition parameters are already available

for each of the participants, then steps S18 and S20 are omitted.

At step S22, central controller 36 causes display
5 processor 64 to display a message on display device 26
requesting the user to perform steps to enable the
camera 2 to be calibrated.

In response, the user carries out the necessary steps
10 and, at step S24, central controller 36 performs
processing to calibrate the camera 2. More particularly,
in this embodiment, the steps performed by the user and
the processing performed by central controller 36 are
carried out in a manner such as that described in
15 Appendix A herewith. This generates calibration data
defining the position and orientation of the camera 2
with respect to the meeting room and also the intrinsic
camera parameters (aspect ratio, focal length, principal
point, and first order radial distortion coefficient).
20 The calibration data is stored in memory 42.

At step S26, central controller 36 causes display
processor 64 to display a message on display device 26
requesting the next participant in the meeting (this
25 being the first participant the first time step S26 is
performed) to sit down.

At step S28, processing apparatus 24 waits for a predetermined period of time to give the requested participant time to sit down, and then, at step S30, central controller 36 processes image data from camera 2 to determine an estimate of the position of the seated participant's head. More particularly, in this embodiment, central controller 36 carries out processing in a conventional manner to identify each portion in a frame of image data from camera 22 which has a colour corresponding to the colour of the skin of the participant (this colour being determined from the data defining the head model of the participant stored in head model store 52), and then selects the portion which corresponds to the highest position in the meeting room (since it is assumed that the head will be the highest skin-coloured part of the body). Using the position of the identified portion in the image and the camera calibration parameters determined at step S24, central controller 36 then determines an estimate of the three-dimensional position of the head in a conventional manner.

At step S32, central controller 36 determines an estimate of the orientation of the participant's head in three dimensions. More particularly, in this embodiment, central controller 36 renders the three-dimensional

computer model of the participant's head stored in head model store 52 for a plurality of different orientations of the model to produce a respective two-dimensional image of the model for each orientation, compares each
5 two-dimensional image of the model with the part of the videoframe from camera 2 which shows the participant's head, and selects the orientation for which the image of the model best matches the video image data. In this embodiment, the computer model of the participant's head
10 is rendered in 108 different orientations to produce image data for comparing with the video data from camera 2. These orientations correspond to 36 rotations of the head model in 10° steps for each of three head inclinations corresponding to 0° (looking straight
15 ahead), +45° (looking up) and -45° (looking down). When comparing the image data produced by rendering the head model with the video data from camera 2, a conventional technique is used, for example as described in "Head Tracking Using a Textured Polygonal Model" by Schödl,
20 Haro & Essa in Proceedings 1998 Workshop on Perceptual User Interfaces.

At step S34, the estimate of the position of the participant's head generated at step S30 and the estimate
25 of the orientation of the participant's head generated at step S32 are input to head tracker 50 and frames of

image data received from camera 2 are processed to track the head of the participant. More particularly, in this embodiment, head tracker 50 performs processing to track the head in a conventional manner, for example as
5 described in "An Analysis/Synthesis Cooperation for Head Tracking and Video Face Cloning" by Valente et al in Proceedings EECV '98 Workshop on Perception of Human Action, University of Freiberg, Germany, June 6 1998.

10 Figure 5 summarises the processing operations performed by head tracker 50 at step S34.

Referring to Figure 5, at step S50, head tracker 50 reads the current estimates of the 3D position and orientation
15 of the participant's head, these being the estimates produced at steps S30 and S32 in Figure 3 the first time step S50 is performed.

At step S52, head tracker 50 uses the camera calibration
20 data generated at step S24 to render the three-dimensional computer model of the participant's head stored in head model store 52 in accordance with the estimates of position and orientation read at step S50.

25 At step S54, head tracker 50 processes the image data for the current frame of video data received from camera 2

to extract the image data from each area which surrounds the expected position of one of the head features identified by the user and stored at step S14, the expected positions being determined from the estimates
5 read at step S50 and the camera calibration data generated at step S24.

At step S56, head tracker 50 matches the rendered image data generated at step S52 and the camera image data
10 extracted at step S54 to find the camera image data which best matches the rendered head model.

At step S58, head tracker 50 uses the camera image data identified at step S56 which best matches the rendered
15 head model to determine the 3D position and orientation of the participant's head for the current frame of video data.

At the same time that step S58 is performed, at step S60,
20 the positions of the head features in the camera image data determined at step S56 are input into a conventional Kalman filter to generate an estimate of the 3D position and orientation of the participant's head for the next frame of video data. Steps S50 to S60 are performed
25 repeatedly for the participant as frames of video data are received from video camera 2.

Referring again to Figure 3, at step S36, central controller 36 determines whether there is another participant in the meeting, and steps S26 to S36 are repeated until processing has been performed for each participant in the manner described above. However, while these steps are performed for each participant, at step S34, head tracker 50 continues to track the head of each participant who has already sat down.

10 When it is determined at step S36 that there are no further participants in the meeting and that accordingly the head of each participant is being tracked by head tracker 50, then, at step S38, central controller 36 causes an audible signal to be output from processing apparatus 24 to indicate that the meeting between the participants can begin.

15

Figure 6 shows the processing operations performed by processing apparatus 24 as the meeting between the participants takes place.

20

Referring to Figure 6, at step S70, head tracker 50 continues to track the head of each participant in the meeting. The processing performed by head tracker 50 at step S70 is the same as that described above with respect to step S34, and accordingly will not be described again

25

here..

At the same time that head tracker 50 is tracking the head of each participant at step S70, at step S72
5 processing is performed to generate and store data in meeting archive database 60.

Figure 7 shows the processing operations performed at step S72.

10

Referring to Figure 7, at step S80, archive processor 58 generates a so-called "viewing parameter" for each participant defining at whom the participant is looking.

15 Figure 8 shows the processing operations performed at step S80.

Referring to Figure 8, at step S110, archive processor 58 reads the current three-dimensional position of each
20 participant's head from head tracker 50, this being the position generated in the processing performed by head tracker 50 at step S58 (Figure 5).

At step S112, archive processor 58 reads the current
25 orientation of the head of the next participant (this being the first participant the first time step S112 is

performed) from head tracker 50. The orientation read at step S112 is the orientation generated in the processing performed by head tracker 50 at step S58 (Figure 5).

5

At step S114, archive processor 58 determines the angle between a ray defining where the participant is looking (a so-called "viewing ray") and each notional line which connects the head of the participant with the centre of
10 the head of another participant.

More particularly, referring to Figures 9 and 10, an example of the processing performed at step S114 is illustrated for one of the participants, namely
15 participant 10 in Figure 1. Referring to Figure 9, the orientation of the participant's head read at step S112 defines a viewing ray 90 from a point between the centre of the participant's eyes which is perpendicular to the participant's head. Similarly, referring to Figure 10,
20 the positions of all of the participant's heads read at step S110 define notional lines 92, 94, 96 from the point between the centre of the eyes of participant 10 to the centre of the heads of each of the other participants 6, 8, 12. At step S114, archive processor 58 determines the
25 angles 98, 100, 102 between the viewing ray 90 and each of the notional lines 92, 94, 96.

Referring again to Figure 8, at step S116, archive processor 58 selects the angle 98, 100 or 102 which has the smallest value. Thus, referring to the example shown in Figure 10, the angle 100 would be selected.

5

At step S118, archive processor 58 determines whether the selected angle has a value less than 10° .

If it is determined at step S118 that the angle is less than 10° , then, at step S120, archive processor 58 sets the viewing parameter for the participant to the number (allocated at step S4 in Figure 3) of the participant connected by the notional line which makes the smallest angle with the viewing ray. Thus, referring to the example shown in Figure 10, if angle 100 is less than 10° , then the viewing parameter would be set to the participant number of participant 6 since angle 100 is the angle between viewing ray 90 and notional line 94 which connects participant 10 to participant 6.

20

On the other hand, if it is determined at step S118 that the smallest angle is not less than 10° , then, at step S122, archive processor 58 sets the value of the viewing parameter for the participant to "0". This indicates that the participant is determined to be looking at none of the other participants since the viewing ray 90 is not

25

close enough to any of the notional lines 92, 94, 96. Such a situation could arise, for example, if the participant was looking at notes or some other object in the meeting room.

5

At step S124, archive processor 58 determines whether there is another participant in the meeting, and steps S112 to S124 are repeated until the processing described above has been carried out for each of the participants.

10

Referring again to Figure 7, at step S82, central controller 36 and voice recognition processor 54 determine whether any speech data has been received from the microphones 4 for the current frame of video data.

15

If it is determined at step S82 that speech data has been received, then, at step S84, archive processor 58 processes the viewing parameters generated at step S80 to determine which of the participants in the meeting is speaking.

20

Figure 11 shows the processing operations performed at step S84 by archive processor 58.

25

Referring to Figure 11, at step S140, the number of occurrences of each viewing parameter value generated at

step S80 is determined, and at step S142, the viewing parameter value with the highest number of occurrences is selected. More particularly, the processing performed at step S80 in Figure 7 will generate one viewing
5 parameter value for the current frame of video data for each participant in the meeting (thus, in the example shown in Figure 1, four values would be generated). Each viewing parameter will have a value which corresponds to the participant number of one of the other participants
10 or "0". Accordingly, at step S140 and S142, archive processor 58 determines which of the viewing parameter values generated at step S80 occurs the highest number of times for the current frame of video data.

15 At step S144, it is determined whether the viewing parameter with the highest number of occurrences has a value of "0" and, if it has, at step S146, the viewing parameter value with the next highest number of occurrences is selected. On the other hand, if it is
20 determined at step S144 that the selected value is not "0", then step S146 is omitted.

At step S148, the participant defined by the selected viewing parameter value (that is, the value selected at
25 step S142 or, if this value is "0" the value selected at step S146) is identified as the participant who is

speaking, since the majority of participants in the meeting will be looking at the speaking participant.

Referring again to Figure 7, at step S86, archive processor 58 stores the viewing parameter value for the speaking participant, that is the viewing parameter value generated at step S80 defining at whom the speaking participant is looking, for subsequent analysis, for example in memory 42.

10

At step S88, archive processor 58 informs voice recognition processor 54 of the identity of the speaking participant determined at step S84. In response, voice recognition processor 54 selects the speech recognition parameters for the speaking participant from speech recognition parameter store 56 and uses the selected parameters to perform speech recognition processing on the received speech data to generate text data corresponding to the words spoken by the speaking participant.

15
20

On the other hand, if it is determined at step S82 that the received sound data does not contain any speech, then steps S84 to S88 are omitted.

25

At step S90, archive processor 58 encodes the current

frame of video data received from camera 2 and the sound data received from microphones 4 as MPEG 2 data in a conventional manner, and stores the encoded data in meeting archive database 60.

5

Figure 12 schematically illustrates the storage of data in meeting archive database 60. The storage structure shown in Figure 12 is notional and is provided for illustration purposes only, to assist understanding; it does not necessarily represent the exact way in which data is stored in meeting archive database 60.

10

Referring to Figure 12, meeting archive database 60 stores time information represented by the horizontal axis 200, on which each unit represents a predetermined amount of time, for example one frame of video data received from camera 2. The MPEG 2 data generated at step S90 is stored as data 202 in meeting archive database 60, together with timing information (this timing information being schematically represented in Figure 12 by the position of the MPEG 2 data 202 along the horizontal axis 200).

15

20

Referring again to Figure 7, at step S92, archive processor 58 stores any text data generated by voice recognition processor 54 at step S88 for the current

25

frame in meeting archive database 60 (indicated at 204 in Figure 12). More particularly, the text data is stored with a link to the corresponding MPEG 2 data, this link being represented in Figure 12 by the text data
5 being stored in the same vertical column as the MPEG 2 data. As will be appreciated, there will not be any text data for storage from participants who are not speaking. In the example shown in Figure 12, text is stored for the first ten time slots for participant 1 (indicated at
10 206), for the twelfth to twentieth time slots for participant 3 (indicated at 208), and for the twenty-first time slot for participant 4 (indicated at 210). No text is stored for participant 2 since, in this example, participant 2 did not speak during the time
15 slots shown in Figure 12.

At step S94, archive processor 58 stores the viewing parameter value generated for each participant at step S80 in the meeting archive database 60 (indicated at 212
20 in Figure 12). Referring to Figure 12, a viewing parameter value is stored for each participant together with a link to the associated MPEG 2 data 202 and the associated text data 204 (this link indicated in Figure 12 by the viewing parameters values being stored in the
25 same column as the associated MPEG 2 data 202 and associated text data 204). Thus, referring to the first

time slot by way of example, the viewing parameter value for participant 1 is "3", indicating that participant 1 is looking at participant 3, the viewing parameter value for participant 2 is "1", indicating that participant 2 is looking at participant 1, the viewing parameter value for participant 3 is also "1", indicating that participant 3 is also looking at participant 1, and the viewing parameter value for participant 4 is "0", indicating that participant 4 is not looking at any of the other participants (in the example shown in Figure 1, the participant indicated at 12 is looking at her notes rather than any of the other participants).

At step S96, central controller 36 and archive processor 58 determine whether one of the participants in the meeting has stopped speaking. In this embodiment, this check is performed by examining the text data 204 to determine whether text data for a given participant was present for the previous time slot, but is not present for the current time slot. If this condition is satisfied for a participant (that is, a participant has stopped speaking), then, at step S98, archive processor 58 processes the viewing parameter values for the participant who has stopped speaking previously stored when step S86 was performed (these viewing parameter values defining at whom the participant was looking

during the period of speech which has now stopped) to generate data defining a viewing histogram. More particularly, the viewing parameter values for the period in which the participant was speaking are processed to generate data defining the percentage of time during that
5 period that the speaking participant was looking at each of the other participants.

Figures 13A and 13B show the viewing histograms
10 corresponding to the periods of text 206 and 208 respectively in Figure 12.

Referring to Figure 12 and Figure 13A, during the period 206 when participant 1 was speaking, he was looking at
15 participant 3 for six of the ten time slots (that is, 60% of the total length of the period for which he was talking), which is indicated at 300 in Figure 13A, and at participant 4 for four of the ten time slots (that is, 40% of the time), which is indicated at 310 in
20 Figure 13A.

Similarly, referring to Figure 12 and Figure 13B, during the period 208, participant 3 was looking at participant 1 for approximately 45% of the time, which is indicated
25 at 320 in Figure 13B, at participant 4 for approximately 33% of the time, indicated at 330 in Figure 13B, and at

participant 2 for approximately 22% of the time, which is indicated at 340 in Figure 13B.

Referring again to Figure 7, at step S100, the viewing
5 histogram generated at step S98 is stored in the meeting
archive database 60 linked to the associated period of
text for which it was generated. Referring to Figure 12,
the stored viewing histograms are indicated at 214, with
the data defining the histogram for the text period 206
10 indicated at 216, and the data defining the histogram for
the text period 208 indicated at 218. In Figure 12, the
link between the viewing histogram and the associated
text is represented by the viewing histogram being stored
in the same columns as the text data.

15 On the other hand, if it is determined at step S96 that,
for the current time period, one of the participants has
not stopped speaking, then steps S98 and S100 are
omitted.

20 At step S102, central controller 36 determines whether
another frame of video data has been received from
camera 2. Steps S80 to S102 are repeatedly performed
while image data is received from camera 2.

25 When data is stored in meeting archive database 60, then

the meeting archive database 60 may be interrogated to retrieve data relating to the meeting.

Figure 14 shows the processing operations performed to
5 search the meeting archive database 60 to retrieve data relating to each part of the meeting which satisfies search criteria specified by a user.

Referring to Figure 14, at step S200, central controller
10 36 causes display processor 64 to display a message on display device 26 requesting the user to enter information defining the search of meeting archive database 60 which is required. More particularly, in this embodiment, central controller 100 causes the
15 display shown in Figure 15A to appear on display device 26.

Referring to Figure 15A, the user is requested to enter information defining the part or parts of the meeting
20 which he wishes to find in the meeting archive database 60. More particularly, in this embodiment, the user is requested to enter information 400 defining a participant who was talking, information 410 comprising one or more key words which were said by the participant identified
25 in information 400, and information 420 defining the participant to whom the participant identified in

information 400 was talking. In addition, the user is able to enter time information defining a portion or portions of the meeting for which the search is to be carried out. More particularly, the user can enter
5 information 430 defining a time in the meeting beyond which the search should be discontinued (that is, the period of the meeting before the specified time should be searched), information 440 defining a time in the meeting after which the search should be carried out, and
10 information 450 and 460 defining a start time and end time respectively between which the search is to be carried out. In this embodiment, information 430, 440, 450 and 460 may be entered either by specifying a time in absolute terms, for example in minutes, or in relative
15 terms by entering a decimal value which indicates a proportion of the total meeting time. For example, entering the value 0.25 as information 430 would restrict the search to the first quarter of the meeting.

20 In this embodiment, the user is not required to enter all of the information 400, 410 and 420 for one search, and instead may omit one or two pieces of this information. If the user enters all of the information 400, 410 and 420, then the search will be carried out to identify each
25 part of the meeting in which the participant identified in information 400 was talking to the participant

identified in information 420 and spoke the key words defined in information 410. On the other hand, if information 410 is omitted, then a search will be carried out to identify each part of the meeting in which the participant defined in information 400 was talking to the participant defined in information 420 irrespective of what was said. If information 410 and 420 is omitted, then a search is carried out to identify each part of the meeting in which the participant defined in information 400 was talking, irrespective of what was said and to whom. If information 400 is omitted, then a search is carried out to identify each part of the meeting in which any of the participants spoke the key words defined in information 410 to the participant defined in information 420. If information 400 and 410 is omitted, then a search is carried out to identify each part of the meeting in which any of the participants spoke to the participant defined in information 420. If information 420 is omitted, then a search is carried out to identify each part of the meeting in which the participant defined in information 400 spoke the key words defined in information 410, irrespective of to whom the key word was spoken. Similarly, if information 400 and 420 is omitted, then a search is carried out to identify each part of the meeting in which the key words identified in information 410 were spoken, irrespective of who said the

key words and to whom.

In addition, the user may enter all of the time information 430, 440, 450 and 460 or may omit one or more
5 pieces of this information.

Once the user has entered all of the required information to define the search, he begins the search by clicking on area 470 using a user input device such as the mouse
10 30.

Referring again to Figure 14, at step S202, the search information entered by the user is read by central controller 36 and the instructed search is carried out.
15 More particularly, in this embodiment, central controller 36 converts any participant names entered in information 400 or 420 to participant numbers using the table 80 (Figure 4), and considers the text information 204 for the participant defined in information 400 (or all
20 participants if information 400 is not entered). If information 420 has been entered by the user, then, for each period of text, central controller 36 checks the data defining the corresponding viewing histogram to determine whether the percentage of viewing time in the
25 histogram for the participant defined in information 420 is equal to or above a threshold which, in this

embodiment, is 25%. In this way, periods of speech (text) are considered to satisfy the criteria that a participant defined in information 400 was talking to the participant defined in information 420 even if the speaking participant looked at other participants while speaking, provided that the speaking participant looked at the participant defined in information 420 for at least 25% of the time of the speech. Thus, a period of speech in which the value of the viewing histogram is equal to or above 25% for two or more participants would be identified if any of these participants were specified in information 420. If the information 410 has been input by the user, then central controller 36 and text searcher 62 search each portion of text previously identified on the basis of information 400 and 420 (or all portions of text if information 400 and 420 was not entered) to identify each portion containing the key word(s) identified in information 410. If any time information has been entered by the user, then the searches described above are restricted to the meeting times defined by those limits.

At step S204, central controller 36 causes display processor 64 to display a list of relevant speeches identified during the search to the user on display device 26. More particularly, central controller 36

causes information such as that shown in Figure 15B to be displayed to the user. Referring to Figure 15B, a list is produced of each speech which satisfies the search parameters, and information is displayed defining the start time for the speech both in absolute terms and as a proportion of the full meeting time. The user is then able to select one of the speeches for playback by clicking on the required speech in the list using the mouse 30.

10

At step S206, central controller 36 reads the selection made by the user at step S204, and plays back the stored MPEG 2 data 202 for the relevant part of the meeting from meeting archive database 60. More particularly, central controller 36 and display processor 64 decode the MPEG 2 data 202 and output the image data and sound via display device 26.

15

At step S208, central controller 36 determines whether the user wishes to cease interrogating the meeting archive database 60 and, if not, steps S200 to S208 are repeated.

20

Various modifications and changes can be made to the embodiment of the invention described above.

25

For example, in the embodiment above the microphones 4 are provided on the meeting room table. However, instead, a microphone on video camera 2 may be used to record sound data.

5

In the embodiment above, image data is processed from a single video camera 2. However, to improve the accuracy with which the head of each participant is tracked, video data from a plurality of video cameras may be processed.

10 For example, image data from a plurality of cameras may be processed as in steps S50 to S56 of Figure 5 and the resulting data from all of the cameras input to a Kalman filter at step S60 in a conventional manner to generate a more accurate estimate of the position and orientation

15 of each participant's head in the next frame of video data from each camera. If multiple cameras are used, then the MPEG 2 data 202 stored in meeting archive database 60 may comprise the video data from all of the cameras and, at steps S204 and S206 in Figure 14 image

20 data from a camera selected by the user may be replayed.

In the embodiment above, the viewing parameter for a given participant defines at which other participant the participant is looking. However, the viewing parameter

25 may also be used to define at which object the participant is looking, for example a display board,

projector screen etc. Thus, when interrogating the meeting archive database 60, information 420 in Figure 15A could be used to specify at whom or at what the participant was looking when he was talking.

5

In the embodiment above, at step S202 (Figure 14), the viewing histogram for a particular portion of text is considered and it is determined that the participant was talking to a further participant if the percentage of gaze time for the further participant in the viewing histogram is equal to or above a predetermined threshold. Instead, however, rather than using a threshold, the participant to whom the speaking participant was looking during the period of text may be defined to be the participant having the highest percentage gaze value in the viewing histogram (for example participant 3 in Figure 13A, and participant 1 in Figure 13B).

In the embodiment above, the MPEG 2 data 202, the text data 204, the viewing parameters 212 and the viewing histograms 214 are stored in meeting archive database 60 in real time as data is received from camera 2 and microphones 4. However, instead, the video and sound data may be stored and data 202, 204, 212 and 214 generated and stored in meeting archive database 60 in non-real-time.

In the embodiment above, the MPEG 2 data 202, the text data 204, the viewing parameters 212 and the viewing histograms 214 are generated and stored in the meeting archive database 60 before the database is interrogated to retrieve data for a defined part of the meeting. However, some, or all, of the data 204, 212 and 214 may be generated in response to a search of the meeting archive database 60 being requested by the user by processing the stored MPEG 2 data 202, rather than being generated and stored prior to such a request. For example, although in the embodiment above the viewing histograms 214 are calculated and stored in real-time at steps S98 and S100 (Figure 7), these histograms could be calculated in response to a search request being input by the user.

In the embodiment above, text data 204 is stored in meeting archive database 60. Instead, audio data may be stored in the meeting archive database 60 instead of the text data 204. The stored audio data would then either itself be searched for key words using voice recognition processing or converted to text using voice recognition processing and the text search using a conventional text searcher.

25

In the embodiment above, processing apparatus 24 includes

functional components for receiving and generating data to be archived (for example, central controller 36, head tracker 50, head model store 52, voice recognition processor 54, speech recognition parameter store 56 and
5 archive processor 58), functional components for storing the archive data (for example meeting archive database 60), and also functional components for searching the database and retrieving information therefrom (for example central controller 36 and text searcher 62).
10 However, these functional components may be provided in separate apparatus. For example, one or more apparatus for generating data to be archived, and one or more apparatus for database searching may be connected to one or more databases via a network, such as the Internet.

15

Also, referring to Figure 16, video and sound data from one or more meetings 500, 510, 520 may be input to a data processing and database storage apparatus 530 (which comprises functional components to generate and store the
20 archive data), and one or more database interrogation apparatus 540, 550 may be connected to the data processing and database storage apparatus 530 for interrogating the database to retrieve information therefrom.

25

In the embodiment above, processing is performed by a

computer using processing routines defined by programming instructions. However, some, or all, of the processing could be performed using hardware.

5 Although the embodiment above is described with respect to a meeting taking place between a number of participants, the invention is not limited to this application, and, instead, can be used for other applications, such as to process image and sound data on
10 a film set etc.

Different combinations of the above modifications are, of course, possible and other changes and modifications can be made without departing from the spirit and scope
15 of the invention.

The contents of the applicant's co-pending UK applications 9905191.4, 9905197.1, 9905202.9, 9905158.3, 9905201.1, 9905186.4, 9905160.9, 9905199.7 and 9905187.2
20 are hereby incorporated by reference.

Calibrating and 3D modelling with a multi-camera system

Charles Wiles and Allan Davison

Computer Vision Group, Canon Research Centre Europe
Guildford, Surrey, UK, GU2 5YJ

Abstract

This paper describes a simple and novel way for calibrating the position and internal camera parameters of a camera viewing a scene with no prior knowledge of the camera being necessary. Only two views of a simple planar grid of spots are used to accurately determine the relative position of each camera in a multiple camera system. A multiple camera system is necessary for modelling dynamic objects (such as people). When the shape of the object is continually changing a large number of images must be taken simultaneously. The multiple camera system is also an important research tool allowing surface generation algorithms to be investigated under known accuracy in the camera positions. We have evaluated our algorithm's performance using simulations to determine the limits on the accuracy of our system and have demonstrated the performance in practice by producing 3D models from a four camera system.

1 Introduction

1.1 Motivation

Computing 3D models of a scene from multiple images observing the scene involves two key steps. First the relative position of the camera to the object being modelled must be determined for each image (*camera solving*), second the 3D structure of the object is computed by intersecting the coloured rays observed in the pixels of each image (*surface generation*).

Various methods exist for computing camera positions. When a single hand-held camera is used to record multiple images of a static scene from different positions the position of the camera can be computed by matching distinguishable *features* on surfaces in the scene between views and employing a *structure from motion* algorithm. Although such an approach works well when the features are accurately matched it can fail when few distinguishable features are visible in the scene. Moreover, such a system fails when the scene is dynamic, containing for example a human being.

To avoid feature matching problems prior to camera

solving the camera positions can be either computed by observing a *calibration* object in the scene or *measured* directly using an alternative device. To model an arbitrary dynamic scene it is necessary to record multiple images from different views at the same instant in time; hence multiple cameras are necessary.

For these reasons, we have explored the use of a calibrated multi-camera 3D modelling system. Not only does such a system allow dynamic objects with few distinguishable features to be modelled, but it provides a valuable research tool for investigating surface generation algorithms, since the accuracy of the camera positions can be independently established. Indeed, the accuracy, coverage and robustness plus the choice of algorithm for surface generation depends greatly on the accuracy of the camera positions determined.

1.2 Issues

There are several important research issues concerning calibrated multiple camera systems:

- Prior knowledge of camera intrinsic parameters
- Ease of production of calibration object
- Ease of calibration process
- The number of images that need to be taken (few images are suitable for still cameras, whereas many images can be used for video cameras).
- Range of camera positions from which the camera can view the calibration object well enough to be calibrated
- Accuracy of calibration
- Accuracy of matching image data to the calibration model

There is clearly a trade off between such factors since the most accurate calibration process would likely require a complicated calibration object and process. However, one of the key aims of our work

has been to find the simplest object and process possible to give us a predefined accuracy in calibration.

We define *object space* to be the limit on 3D space within which the object to be modelled is assumed to lie (for a human being this might be a 2x2x2m cube of space). Our goal in calibrating the camera positions is to recover them such that any point within object space projects to within 1 pixel of its true projection in the image. By achieving *maximum projection error* of less than 1 pixel we limit the range of search for agreeing texture in image co-ordinates to 1 pixel from predicted positions during surface generation.

Unfortunately this does not give us a clear measure of the accuracy of the surface generated when correct matches are found between images under surface generation. It does, however, guarantee that the surface of the model will project to within 1 pixel of its true observed position in the original images. We argue that this measure of accuracy on camera position is more important when the goal is to generate a model that is *photo-consistent* with the original images.

1.3 Background

The most accurate calibration object and process would be to have a known 3D point observed in the image for every point in object space and to compute the transfer equation that projects these points into the observed image co-ordinates.

In practice the act of projection is assumed to be a simple parameterised type called a *camera model*. If the parameters (known as the *intrinsic parameters*) of this projection are known then only three world-to-image point matches are required in order to fix the six degrees of freedom in the unknown orientation and location of the camera (known as the camera *extrinsic parameters*)¹. If the intrinsic parameters are not known then, depending on the nature of the calibration object, it is possible to compute these parameters at the same time as the extrinsic parameters from a larger number of matches². In practice noise in the image measurements of the observed points is inevitable and many matches are required so that the maximum likelihood solution can be found by least squares.

In our work we have assumed that the camera model may radially distort the image during projection and that the intrinsic parameters are unknown in advance. Given this starting point the problem is significantly more complicated than computing just

the extrinsic parameters when the intrinsic parameters are known in advance. One reason for taking this approach is that although manufacturers often provide accurate values for the *focal length* of their cameras, the position of the *principal point* often varies greatly and is unknown. Accurate knowledge of the principal point is vital for computing accurate camera position from coplanar world points (see [4]).

There are several methods that have been used to calibrate both intrinsic and extrinsic parameters of a camera. Perhaps the simplest method is to take a single image of a planar calibration grid of known structure. This method was pioneered by Tsai [7]. Although this method is simple, it is only capable of completely calibrating both the intrinsic and extrinsic parameters of the camera if the imaging process exhibits significant radial distortion. If there is little or no radial distortion, the position of the principal point in the image cannot be determined independently from the height of the camera above the calibration plane, and hence calibration fails to provide a complete answer.

In order to calibrate the intrinsic parameters from a planar grid when no radial distortion occurs, multiple images of the grid must be taken. This is the method used by Kanade et al [2] to calibrate their multi-camera rig for dynamic 3D modelling of people. First a planar grid is moved randomly around in front of each video camera to calibrate the intrinsic parameters of the camera, second each camera's position is computed relative to a grid of spots on the floor. The method gives good accuracy of calibration, but uses many images for each camera to carry out the calibration. Our method, presented later, adapts this method to work with just two images from a still camera.

The problem with a single image of a planar grid is solved if an accurate three-dimensional calibration grid is manufactured. Typically an "L" shaped grid is used which has a pattern of black squares on a white background on each of the two flat surfaces. Such calibration grids allow reasonably accurate calibration of the camera to be performed from a single image of the grid and this method has been used extensively for calibrating stereo-camera rigs [1]. The main drawback of such an approach is that accurate manufacturing of the calibration grid is necessary and this is both awkward and expensive. Moreover, such a system does not scale well to large environments and the calibration object must be carefully orientated so that all cameras see a "good" view of the object.

An alternative method is to use a planar calibration

¹In fact, exactly four different solutions are obtained under projective imaging conditions and exactly two different solutions are obtained under affine imaging conditions.

²For example for the perspective camera model, the four intrinsic and six extrinsic parameters can be computed from five 2D-3D point matches when the points are non-coplanar.

grid, but to move it through space in a *known* motion. Either a robot arm or a stepper motor is used to move the grid. By moving the grid in known steps and taking an image after each step, the whole object space can be swept out providing a wealth of matches. Such systems can provide very accurate camera calibration, and although the planar calibration object can be easily made the main drawback is in the expense of the robot arm or stepper motor and in the lack of scalability of the system. Since many images should be recorded for each camera, it is suitable for calibrating video cameras, but is not an ideal method for calibrating still cameras.

More recently "magic wands" [10] have been used to calibrate multiple camera systems. The magic wand approach is to move a single point (or pair of points one at each end of a wand) by hand randomly around in space. Typically the scene is darkened and a point light source is used. The actual algorithm for computing the extrinsic parameters is similar to self-calibration [6] from an unknown object (or structure from motion, if the intrinsic parameters of the cameras are known), but is simpler since the matching problem is trivially solved between images. This method scales well and is inexpensive. It is an excellent method for calibrating multiple video camera rigs, but due to only one point appearing in each image it is not suitable for the calibration of multiple still camera rigs. A further consideration for the magic wand approach are that if a full self-calibration technique is used with a wand with a single point certain camera configurations must be avoided (for example, the cameras must not all lie in the same plane). However if a wand with points at both ends is used then the ambiguity in the Euclidean reconstruction is removed and full self-calibration is possible.

2 Accurate calibration from a planar grid

Our method for calibrating a multiple still camera rig uses a simple planar calibration grid of regularly spaced black circles on a white background. Such a grid is trivial to manufacture (by for example printing on a standard home printer).

The challenge then is to accurately calibrate the position and intrinsic parameters of a number of cameras that observe the grid. Our method stems from the observation that all the intrinsic parameters of a camera can be accurately determined from a small number of images of a planar grid taken with the camera at different positions without needing to know any of the camera positions in advance. Indeed if the camera positions are chosen reasonably carefully, the calibration

can be done from just two images. If one of these images was recorded with the camera in its final position in the multi-camera system then accurate calibration of the multi-camera system can be achieved with just two images from each camera. Additional images provide greater accuracy to the calibration.

The process for calibration is as follows:

1. Take an image of the calibration grid with the camera rotated by roughly 90° about the viewing direction and with a different tilt from the final orientation in the multiple camera rig.
2. Put the camera in its final position in the multiple camera rig and secure in place. Take a second image of the calibration grid.

With the camera in its final position, the calibration grid can be removed and objects to be modelled placed within the space observed by the multi-camera rig.

2.1 Calibrating from a single image

Tsai [7] pioneered the process of calibrating a camera from a planar grid. Tsai's camera model projecting world point $\mathbf{X} = (X, Y, Z)^T$ to image point $\mathbf{x} = (x, y)^T$ is defined by the set of equations:

$$\mathbf{x} = 1/(1 + \kappa_1 R_n^2) \begin{bmatrix} \xi f & 0 \\ 0 & f \end{bmatrix} \mathbf{x}_n + \mathbf{x}_0,$$

where

$$R_n^2 = x_n^2 + y_n^2$$

and

$$\mathbf{x}_n = \frac{1}{Z_c} \begin{pmatrix} X_c \\ Y_c \end{pmatrix}, \quad \text{where } \mathbf{X}_c = \mathbf{R}\mathbf{X} + \mathbf{t}.$$

\mathbf{R} and \mathbf{t} are the 3×3 rotation matrix and translation vector representing the position of the camera, and aspect ratio ξ , focal length f , principal point $\mathbf{x}_0 = (x_0, y_0)^T$ and first order radial distortion coefficient κ_1 are the five intrinsic parameters of the camera. If the aspect ratio is known and the distortion coefficient is significant then all the intrinsic and extrinsic parameters can be computed from five or more world-to-image point matches. As well as the radial distortion coefficient being significant it is important that the projection of the grid in the image exhibits significant perspective effects. Hence the camera cannot be calibrated if it is far from the calibration grid or if the viewing direction is close to perpendicular to the plane of the grid. In practice this is easy to avoid.

However, when the first order radial distortion coefficient is zero then the imaging equation reduces to the standard, linear perspective camera model which

has 10 parameters (4 intrinsic + 6 extrinsic). In this case the planar grid to image mapping is completely defined by a linear planar homography containing 8 independent parameters. Hence even when the aspect ratio is known the total number of parameters to be estimated in order to calibrate the camera is still 9 (3 unknown intrinsic + 6 extrinsic) which is clearly not possible.

In the rest of this section we argue that the position of the principal point in the direction parallel to the plane of the calibration grid is determined uniquely, but that the component perpendicular to the plane of the calibration grid cannot be determined independently from the rest of the camera parameters. Indeed we gain an insight into this problem by considering the case when the x -axis of the camera is known to be parallel to the XZ -plane of the world so that the roll of the camera is zero with respect to this plane.

The equation for the perspective camera model in homogeneous co-ordinates is

$$\begin{pmatrix} x \\ 1 \end{pmatrix} = P \begin{pmatrix} X \\ 1 \end{pmatrix},$$

where

$$P = K [R \quad t] \quad , \quad K = \begin{bmatrix} \xi f & 0 & x_0 \\ 0 & f & y_0 \\ 0 & 0 & 1 \end{bmatrix}$$

and equality is only defined up to an arbitrary scale factor. In this case of no camera roll we can rewrite the perspective camera model in terms of only the pitch, α , and yaw, β , of the rotation matrix:

$$R = \begin{bmatrix} 1 & 0 & 0 \\ 0 & c_\alpha & s_\alpha \\ 0 & -s_\alpha & c_\alpha \end{bmatrix} \begin{bmatrix} c_\beta & 0 & s_\beta \\ 0 & 1 & 0 \\ -s_\beta & 0 & c_\beta \end{bmatrix}$$

where c_θ and s_θ are the cosine and sine of angle θ respectively.

If we further note that under calibration from a planar grid, the world points on the grid are considered to be in the world plane $Y = 0$, then we can remove the second column of the rotation matrix, R_2 , and the Y structure co-ordinate from the equation for projection. Thus we can rewrite the equation for projection in terms of the 3×3 planar homography H ,

$$\begin{pmatrix} x \\ 1 \end{pmatrix} = K [R_1 \quad R_3 \quad t] \begin{pmatrix} X \\ Z \\ 1 \end{pmatrix} = H \begin{pmatrix} X \\ Z \\ 1 \end{pmatrix}.$$

Simplification of H leads to,

$$\begin{pmatrix} x \\ 1 \end{pmatrix} = \begin{bmatrix} f_g c_\beta - x_0 s_\beta & f_g s_\beta - x_0 c_\beta & t'_x \\ y_h (-s_\beta) & y_h c_\beta & t'_y \\ -s_\beta & c_\beta & t'_z \end{bmatrix} \begin{bmatrix} X \\ Z \\ 1 \end{bmatrix}$$

where $f_g = \xi f / c_\alpha$, $y_h = y_0 + f \tan(\alpha)$ is the *horizon* of the calibration plane in the image and t' contains some reparameterisation of the vector t .

The important observation is that only seven independent parameters can be observed in H despite having nine elements in the matrix (we have fixed the scale using the bottom row of the matrix). Having fixed the roll angle to be zero and assuming that the aspect ratio is known, we would like to compute all eight unknown perspective camera parameters (3 unknown intrinsic + 5 unknown extrinsic). However of these only the yaw, and the x co-ordinate of the principal point can actually be computed.

Although the argument above is valid for only special case of zero camera roll we have observed empirically that for arbitrary camera orientation, the roll, yaw and component of the principal point parallel to the plane of the calibration grid can indeed be observed from a single image for arbitrary camera roll. The pitch, focal length, translation vector and component of the principal point perpendicular to the plane of the calibration grid cannot be observed.

2.2 Calibrating from a pair of images

The observation that for a perspective camera one component of the principal point can be determined from a single view, but that the perpendicular component cannot, leads directly to a simple method for calibration from two views of the calibration grid.

With two views of the grid, assuming the intrinsic parameters of the camera are unchanged between views, there are 16 independent parameters determined by the two homographies and there are 16 parameters to be estimated in the two cameras (4 fixed intrinsic parameters and 2 sets of 6 extrinsic parameters). With known aspect ratio there are only 15 parameters to be estimated, but in either case there are theoretically enough equations to determine all the unknown parameters.

The key question then is, given that there are theoretically enough equations to solve for all the unknowns, *under what conditions is this possible?*

Recall that in the special case of zero roll we are left with two equations in the unobservable parameters:

$$f_g = \xi f / c_\alpha \quad \text{and} \quad y_h = y_0 + f \tan(\alpha).$$

We will consider that we require only to determine the four remaining unknown parameters in these two equations having solved from a single view for the other intrinsic and extrinsic parameters. Each additional view of the calibration grid gives one more set of these two equations but introduces one additional unknown (the pitch of the camera in the new view).

For the case of known aspect ratio and two views then the equations above give four equations in four remaining unknown parameters. Examination of the equations above show that as long as the pitch of the camera changes between the two views of the calibration grid, then the camera can be fully calibrated. Our first conjecture is then

- if the aspect ratio is known in advance then as long as the pitch of the camera changes between two views then the camera can be fully calibrated from two views of a planar calibration grid.

For the case of unknown aspect ratio there are four equations in five remaining unknown parameters and four equations. Hence calibration cannot be computed from two views of the grid. Our second conjecture is then

- if the aspect ratio is not known in advance and there is no change in the roll angle between the two views then the camera cannot be fully calibrated from two views of a planar calibration grid.

However, for the case of when there is a change in the roll angle of the camera between two views then an interesting simplification occurs, since the principal point becomes uniquely determined. This is because each view fixes the principal point to lie on a specific line in the image, the line being perpendicular to the plane of the calibration grid. If the roll angle changes these lines will not be parallel and hence will intersect at the location of the principal point. Algebraically, y_0 is effectively known in the above equations reducing the number of unknown parameters from five to four. So our third and final conjecture is that:

- if the aspect ratio is not known in advance and there is both a change in the roll angle and a change in the tilt angle between the two views then the camera can be fully calibrated from two views of a planar calibration grid.

Hence in order for the camera intrinsic parameters to be determined fully from two views there must be relative roll (with respect to the plane of the calibration grid) between the positions of the two cameras. Indeed ideally the relative roll should be 90° .

2.3 Accurate feature location

Our approach to matching the calibration object to image data has been to extract the location in the image of point features of known location on the calibration object forming world-to-image point matches. These co-ordinates are then fed into an algorithm for calibrating the camera parameters. A key goal of our

work is to determine the accuracy in feature location needed in order to guarantee a maximum projection error in object space of less than 1 pixel.

More accurate calibration may well be achievable by following the initial "calibration from matches" stage with an iterative "template matching" stage. This second stage would aim to globally minimise the difference between a template of the calibration object and the raw image data. We have not yet implemented such a stage and hence our evaluation for calibration performance is based on the first stage only and may be improved by template matching.

3 Results

3.1 Synthetic data

In order to validate experimentally our prediction that only one component of the principal point can be localised accurately under calibration from a single plane we performed the following experiment using synthetic data.

First the intrinsic parameters and position of a camera were computed fully automatically from a typical view of a calibration grid with the aspect ratio known in advance. The roll of the camera with respect to the plane of the calibration grid was zero and the camera was tilted so that the centre of the grid projects to the centre of the image. The grid fills the unit square and is centred at (0.5, 0.5, 0.0) in the world. The camera is at (0.7, -1.0, 1.0). Since the roll of the camera is zero we expect that the x co-ordinate of the principal point will be accurately located and the y co-ordinate less accurately located getting progressively worse as the amount of radial distortion decreases.

We then repeated each experiment using two views of the grid, one view as in the single view experiment and a second view rotated by 90° about the viewing direction. The aspect ratio was calculated being assumed unknown in advance.

In both cases the maximum projection error within the unit cube of space above the calibration grid was measured as well as the projection error of the point in the centre of the unit cube.

Tsai's freely available code was used to carry out the minimisation with slight modification to enable minimisation of the two view case. Note that for the case of a single view with no radial distortion the minimisation is clearly under-constrained and hence it would be wise reparameterise the solution. However we found this to be unnecessary since the full minimisation reliably and rapidly converged on a solution in which the observable parameters were accurately recovered.

The process for solving for the two view case was as follows:

1. Solve for each view separately using Tsai's single view approach (using a fixed prior estimate for the aspect ratio).
2. Determine the roll of the camera.
3. From the principal point for each view extract the component parallel to the calibration plane and combine into an initial estimate of the principal point.
4. Solve again for each view separately keeping this principal point fixed.
5. Average the intrinsic parameters from each view and use these and the two sets of extrinsic parameters as an initial estimate for a final full two view minimisation.

These experiments were then repeated with non zero roll. The solution computed from a single view accurately recovers the roll of the camera in the extrinsic parameters and hence it is straight forward to extract and combine the components of the principal point parallel to the calibration plane as a precursor to full two view minimisation.

3.1.1 Effect of varying RMS image noise on errors

In this experiment the performance of the full calibration method was measured as the noise in the image co-ordinates of the matches was varied. The experiment was repeated with varying amounts of radial distortion. The full results are can be found in [8].

The main observations for calibration from a single view were:

- The estimation of the y co-ordinate of the principal point, y_0 , is computed less accurately than the x co-ordinate of the principal point, x_0 .
- The accuracy of the y co-ordinate of the principal point decreases as the amount of radial distortion decreases.
- The accuracy of the x co-ordinate of the principal point is independent of the amount of radial distortion.
- The maximum projection error of the unit cube sitting directly above the calibration grid is highly correlated with the error in y_0 .

- The maximum projection error occurs for points farthest from the calibration grid and is approximately ten times greater than the error in the projection of the centre point of the unit cube.

In essence when the roll of the camera is zero the estimation of the y_0 intrinsic parameter from a single view is poorly constrained. This leads to errors in projection of 3D points and the further away from the plane of the calibration grid the 3D point being projected is the greater the error in the projection. However, the x_0 intrinsic parameter is well constrained.

The main observations from two views were:

- The error in the y_0 intrinsic parameter was recovered to the same level of accuracy as the x_0 parameter irrespective of the amount of radial distortion
- The maximum projection error for both views was greatly reduced compared with the projection error observed from a single view with the same value of image noise. This difference became more marked the lower the amount of radial distortion.
- With no radial distortion in order to calibrate the camera so that the maximum projection error is guaranteed to be less than one pixel the RMS image noise in the image co-ordinates of features must be less than 0.05 pixels.
- With no radial distortion in order to guarantee the centre projection error to be less than one pixel the image noise must be less than 0.25 pixels.

3.2 Real data

Two experiments were carried out, both using the same set of four off-the-shelf PowershotA5 cameras. In the first experiment, the cameras were positioned around a toy dinosaur to demonstrate small scale modelling, whereas in the second experiment the cameras were positioned in a room so that a person could be modelled.

3.3 Toy dinosaur

A calibration grid printed onto a sheet of A4 paper was placed on a stand. Each of the four cameras were held in a portrait orientation and a photograph of the grid taken as shown in the first column of Figure 2.

Then the cameras were set up as shown in Figure 1 with each camera in a landscape orientation. Three cameras were positioned roughly by hand so that they were about 20cm above the plane of the calibration grid and evenly spaced in a circle about the grid. The fourth camera was positioned so that it was roughly

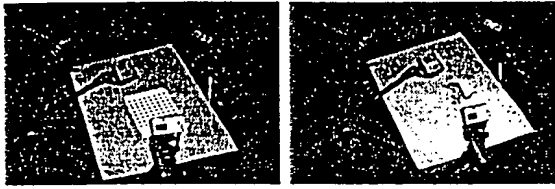


Figure 1: Camera configuration. a) calibration grid imaged, b) toy dinosaur imaged.

above the centre of the grid looking down. A second photograph of the grid was taken with each camera as shown in the second column of Figure 2.

The grid was then removed and in keeping with the indomitable spirit of the academic computer vision community a toy dinosaur placed in the space where the grid had been. A photograph of the dinosaur was taken with each camera.

The three images from each camera were then downloaded to a PC. Two frame calibration as described above was carried out for each camera in turn to calculate its intrinsic parameters and position with respect to the calibration grid in the second image.

Finally the dinosaur images were segmented from the background using a blue screening technique and a voxel carve applied to work out an outer bound on the space occupied by the toy dinosaur. The resulting voxelisation was transformed using a marching cubes algorithm into a faceted VRML model for display. Figure 3 shows the resulting model viewed from the same direction as given in one of the images. The accuracy of the camera calibration is demonstrated by a plausible reconstruction of the toy dinosaur. In particular the tail of the dinosaur is well reconstructed.

3.3.1 Person

A calibration grid was made by sticking together 63 sheets of A4 paper each with a single black circle printed on each. The 7x9 grid occupied a space approximately 1.5m x 1.5m.

First a photograph of the grid was taken with each camera in a landscape orientation. Then the cameras were placed in portrait orientation so that they were roughly evenly spaced through 180°. A second photograph was taken of the calibration grid as shown in the first row of Figure 4. Finally the calibration grid was removed from the scene and a photograph of a person taken with each camera.

A faceted model was reconstructed as before as

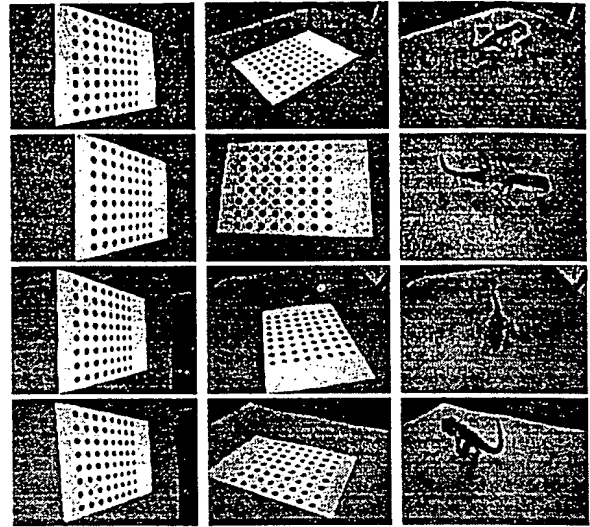


Figure 2: Toy dinosaur images. First column: photographs taken of grid with each camera in portrait orientation (roll angle approximately 90°). Second column: photographs taken with camera in final position in landscape orientation (roll angle approximately 0°). Third column: photographs of toy dinosaur with cameras in final position.



Figure 3: View of the 3D model computed from the toy dinosaur images

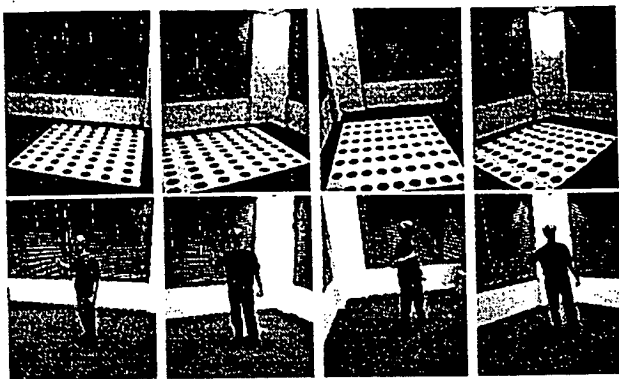


Figure 4: Person images. First row: photographs of the calibration grid with cameras in final position. Second row: photographs of a person.

shown in Figure 5. The accuracy of the camera calibration is demonstrated by a plausible reconstruction of the person.

4 Conclusion

We have shown that accurate camera calibration can be achieved with a simple two views of a plane technique and have demonstrated its practicality by using it to calibrate a multi-camera system for modelling real objects of varying size. Future work will focus on improved techniques for surface generation.

Acknowledgments

The authors would like to thank Ricahrd Taylor, Jane Haslam, Adam Baumberg, Alex Lyons, Simon Rowe and Mike Taylor for their software implementations and Philip McLauchlan for helpful discussions.

References

- [1] O. Faugeras. *Three-Dimensional Computer Vision*. The MIT Press. 1993.
- [2] T. Kanade, P.J. Narayanan, and P.W. Rander. "Virtualized Reality: Constructing virtual worlds from real scenes". *IEEE Multimedia*, 4(1), May 1997.
- [3] A. Katkere, S. Moezzi, D.Y. Kuramura, P. Kelly and R. Jain. "Towards video-based immersive environments". *Multimedia Systems*. 5:69-85, 1997.
- [4] R.K. Lenz, and R.Y. Tsai. "Techniques for calibration of the scale factor and image centre for high accuracy 3D machine vision metrology". *Proc. IEEE Int. Conf. Robotics and Automation*, Raleigh, NC, 68-75, March 1987.
- [5] W. Niem and J. Wingbermühle. "Automatic reconstruction using a mobile monoscopic camera". *Proc. International Conference on Recent Advances in 3D Imaging and Modelling*. Ottawa, Canada, 12-15 May 1997.
- [6] M. Pollefeys, R Koch and L Van Gool. "Self-calibration and metric reconstruction in spite of varying and unknown intrinsic camera parameters". *Proc. International Conference on Computer Vision*, Bombay, India, January 1998.
- [7] R.Y. Tsai. "A versatile camera calibration technique for high accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses". *IEEE Journal of Robotics and Automation*, RA-3(4):323-344, 1987.
- [8] C. Wiles. "Calibrating and 3D modelling with a multi-camera system". Canon Technical Report CRE-TR-98-043, 14 December 1998.
- [9] Geometrix Inc. <http://www.geometrixinc.com/>
- [10] Oxford Metrics. <http://www.oxfordmetrics.co.uk/>
- [11] Vanguard. <http://www.robots.ox.ac.uk/~vanguard/>



Figure 5: View of the 3D model computed from person images

CLAIMS

1. Apparatus for processing image data and sound data, comprising:

5 image processing means for processing image data recorded by at least one camera showing the movements of a plurality of people to determine where each person is looking and to determine which of the people is speaking based on where the people are looking; and

10 sound processing means for processing sound data defining words spoken by the people to generate text data therefrom in dependence upon the result of the processing performed by the image processing means.

15 2. Apparatus according to claim 1, wherein the sound processing means includes storage means for storing respective voice recognition parameters for each of the people, and means for selecting the voice recognition parameters to be used to process the sound data in
20 dependence upon the person determined to be speaking by the image processing means.

3. Apparatus according to claim 1 or claim 2, wherein the image processing means is arranged to determine where
25 each person is looking by processing the image data using camera calibration data defining the position and

orientation of each camera from which image data is processed.

4. Apparatus according to any preceding claim, wherein
5 the image processing means is arranged to determine where each person is looking by processing the image data to track the position and orientation of each person's head in three dimensions.
- 10 5. Apparatus according to any preceding claim, wherein the image processing means is arranged to determine which person is speaking based on the number of people looking at each person.
- 15 6. Apparatus according to claim 5, wherein the image processing means is arranged to generate a value for each person defining at whom the person is looking and to process the values to determine the person who is speaking.
- 20 7. Apparatus according to any preceding claim, wherein the image processing means is arranged to determine that the person who is speaking is the person at whom the most other people are looking.
- 25 8. Apparatus according to any preceding claim, further

comprising a database for storing the image data, the sound data, the text data produced by the sound processing means and viewing data defining where each person is looking, the database being arranged to store the data such that corresponding text data and viewing data are associated with each other and with the corresponding image data and sound data.

9. Apparatus according to claim 8, further comprising means for compressing the image data and the sound data for storage in the database.

10. Apparatus according to claim 9, wherein the means for compressing the image data and the sound data comprises means for encoding the image data and the sound data as MPEG data.

11. Apparatus according to any of claims 8 to 10, further comprising means for generating data defining, for a predetermined period, the proportion of time spent by a given person looking at each of the other people during the predetermined period, and wherein the database is arranged to store the data so that it is associated with the corresponding image data, sound data, text data and viewing data.

12. Apparatus according to claim 11, wherein the predetermined period comprises a period during which the given person was talking.

5 13. Apparatus for processing image data, comprising image processing means for processing image data recorded by at least one camera showing the movements of a plurality of people to determine where each person is looking and to determine which of the people is speaking
10 based on where the people are looking.

14. Apparatus according to claim 13, wherein the image processing means is arranged to determine where each person is looking by processing the image data using
15 camera calibration data defining the position and orientation of each camera from which image data is processed.

15. Apparatus according to claim 13 or claim 14, wherein
20 the image processing means is arranged to determine where each person is looking by processing the image data to track the position and orientation of each person's head in three dimensions.

25 16. Apparatus according to any of claims 13 to 15, wherein the image processing means is arranged to

determine which person is speaking based on the number of people looking at each person.

17. Apparatus according to claim 16, wherein the image
5 processing means is arranged to generate a value for each person defining at whom the person is looking and to process the values to determine the person who is speaking.

10 18. Apparatus according to any of claims 13 to 17, wherein the image processing means is arranged to determine that the person who is speaking is the person at whom the most other people are looking.

15 19. A method of processing image data and sound data, comprising:

an image processing step of processing image data recorded by at least one camera showing the movements of a plurality of people to determine where each person is
20 looking and to determine which of the people is speaking based on where the people are looking; and

a sound processing step of processing sound data defining words spoken by the people to generate text data therefrom in dependence upon the result of the processing
25 performed in the image processing step.

20. A method according to claim 19, wherein the sound processing step includes selecting, from stored respective voice recognition parameters for each of the people, the voice recognition parameters to be used to
5 process the sound data in dependence upon the person determined to be speaking in the image processing step.

21. A method according to claim 19 or claim 20, wherein, in the image processing step, it is determined where each
10 person is looking by processing the image data using camera calibration data defining the position and orientation of each camera from which image data is processed.

15 22. A method according to any of claims 19 to 21, wherein, in the image processing step, it is determined where each person is looking by processing the image data to track the position and orientation of each person's head in three dimensions.

20

23. A method according to any of claims 19 to 22, wherein, in the image processing step, it is determined which person is speaking based on the number of people looking at each person.

25

24. A method according to claim 23, wherein, in the

image processing step, a value is generated for each person defining at whom the person is looking and the values are processed to determine the person who is speaking.

5

25. A method according to any of claims 19 to 24, wherein, in the image processing step, it is determined that the person who is speaking is the person at whom the most other people are looking.

10

26. A method according to any preceding claim, further comprising the step of storing the image data, the sound data, the text data produced by in sound processing step and viewing data defining where each person is looking in a database, the database being arranged to store the data such that corresponding text data and viewing data are associated with each other and with the corresponding image data and sound data.

20 27. A method according to claim 26, wherein the image data and the sound data are stored in the database in compressed form.

28. A method according to claim 27, wherein the image data and the sound data are stored as MPEG data.

25

29. A method according to any of claims 26 to 28, further comprising the steps of generating data defining, for a predetermined period, the proportion of time spent by a given person looking at each of the other people
5 during the predetermined period, and storing the data in the database so that it is associated with the corresponding image data, sound data, text data and viewing data.

10 30. A method according to claim 29, wherein the predetermined period comprises a period during which the given person was talking.

31. A method according to any of claims 26 to 30,
15 further comprising the step of generating a signal conveying the database with data therein.

32. A method according to claim 31, further comprising the step of recording the signal either directly or
20 indirectly to generate a recording thereof.

33. A method of processing image data, comprising processing image data recorded by at least one camera showing the movements of a plurality of people to
25 determine where each person is looking and to determine which of the people is speaking based on where the people

are looking.

34. A method according to claim 33, wherein it is determined where each person is looking by processing the
5 image data using camera calibration data defining the position and orientation of each camera from which image data is processed.

35. A method according to claim 33 or claim 34, wherein
10 it is determined where each person is looking by processing the image data to track the position and orientation of each person's head in three dimensions.

36. A method according to any of claims 33 to 35,
15 wherein it is determined which person is speaking based on the number of people looking at each person.

37. A method according to claim 36, wherein a value is generated for each person defining at whom the person is
20 looking and the values are processed to determine the person who is speaking.

38. A method according to any of claims 33 to 37,
wherein it is determined that the person who is speaking
25 is the person at whom the most other people are looking.

39. A storage device storing instructions for causing a programmable processing apparatus to become configured as an apparatus as set out in any of claims 1 to 18.

5 40. A storage device storing instructions for causing a programmable processing apparatus to become operable to perform a method as set out in any of claims 19 to 38.

10 41. A signal conveying instructions for causing a programmable processing apparatus to become configured as an apparatus as set out in any of claims 1 to 18.

15 42. A signal conveying instructions for causing a programmable processing apparatus to become operable to perform a method as set out in any of claims 19 to 38.

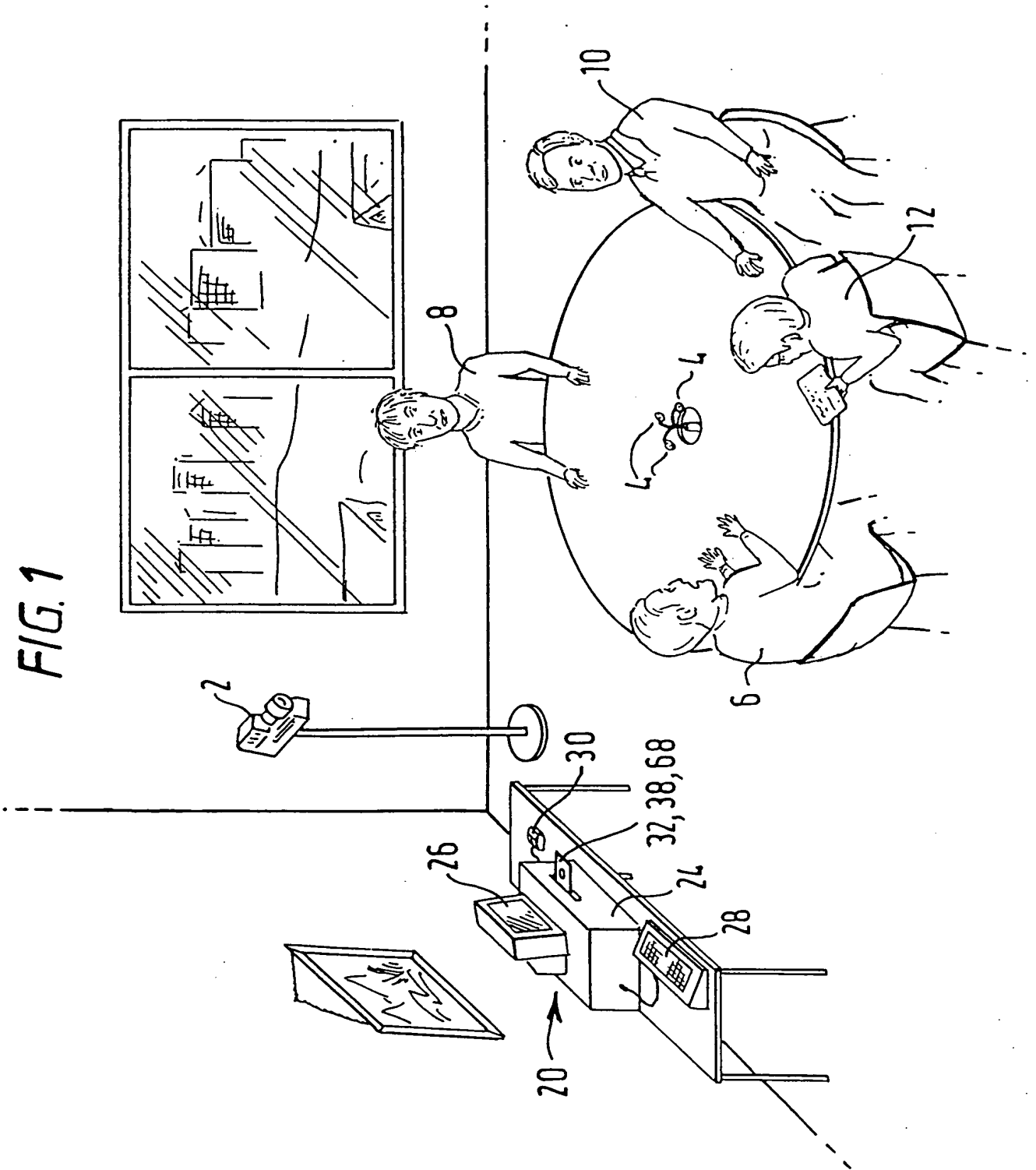
ABSTRACTIMAGE PROCESSING APPARATUS

Image data showing the movements of a number of people,
5 for example in a meeting, and sound data defining the
words spoken by the people is processed by a computer
processing apparatus 24 to archive the data in a meeting
archive database 60. The image data is processed to
10 determine the three-dimensional position and orientation
of each person's head and to determine at whom each
person is looking. Processing is carried out to
determine who is speaking by determining at which person
most people are looking. Having determined which person
15 is speaking, the personal speech recognition parameters
for that person are selected and used to convert the
sound data to text data. The image data, sound data,
text data and data defining at whom each person is
looking is stored in the meeting archive database 60.

20 (FIGURE 2)

THIS PAGE BLANK (USPTO)

FIG. 1



THIS PAGE BLANK (USPTO)

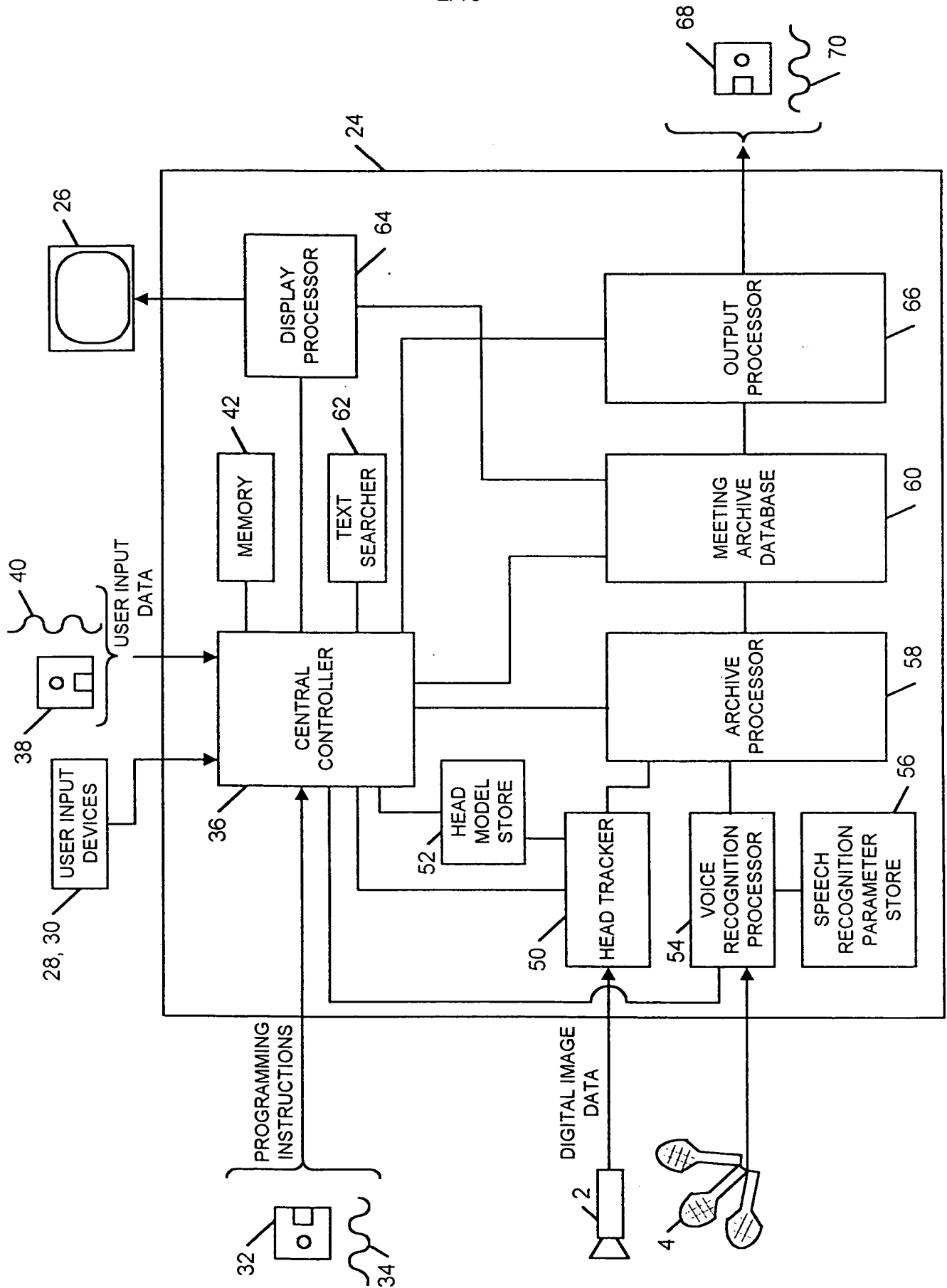
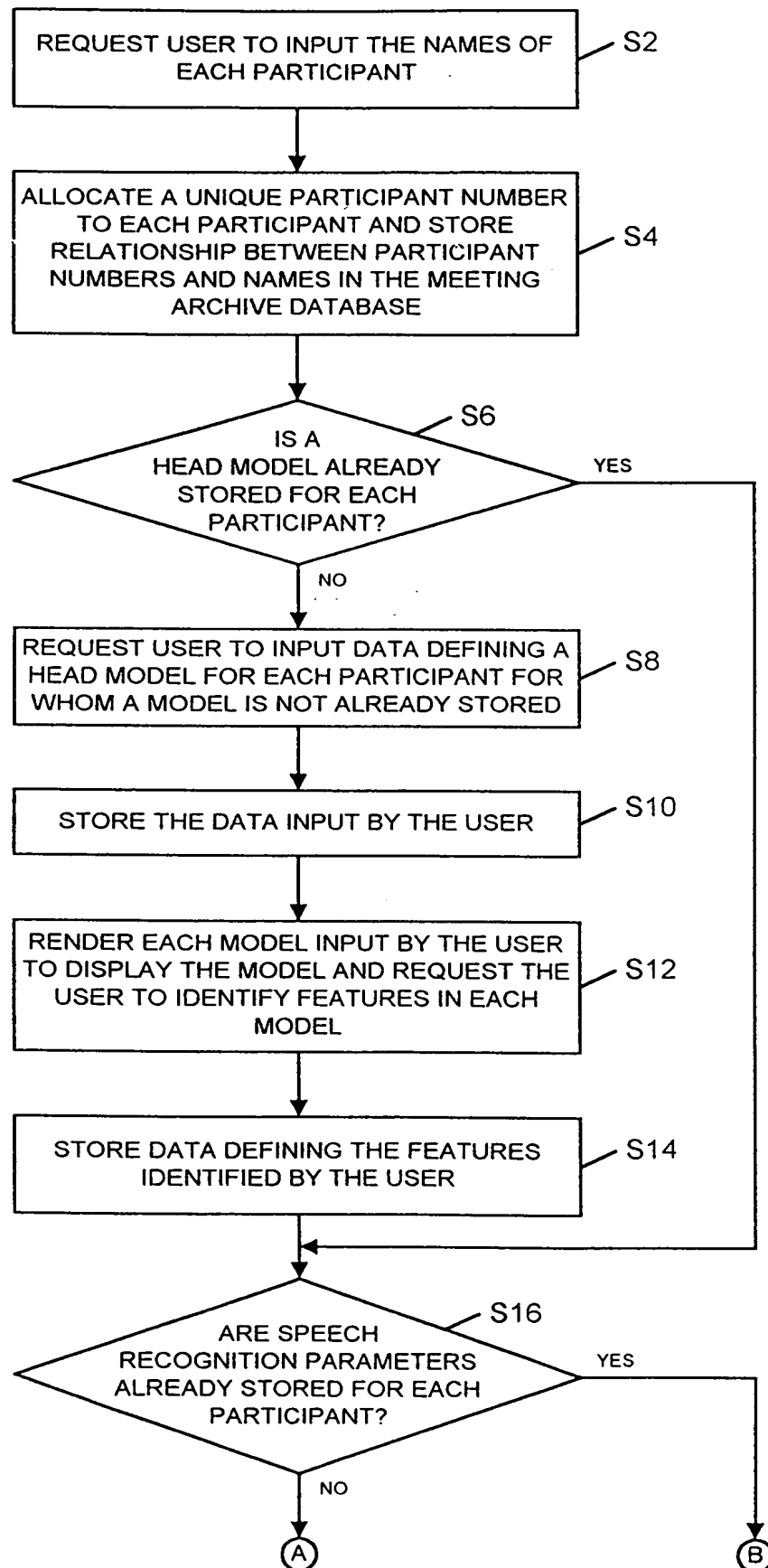


FIG. 2

THIS PAGE BLANK (USPTO)

FIG. 3



THIS PAGE BLANK (USPTO)

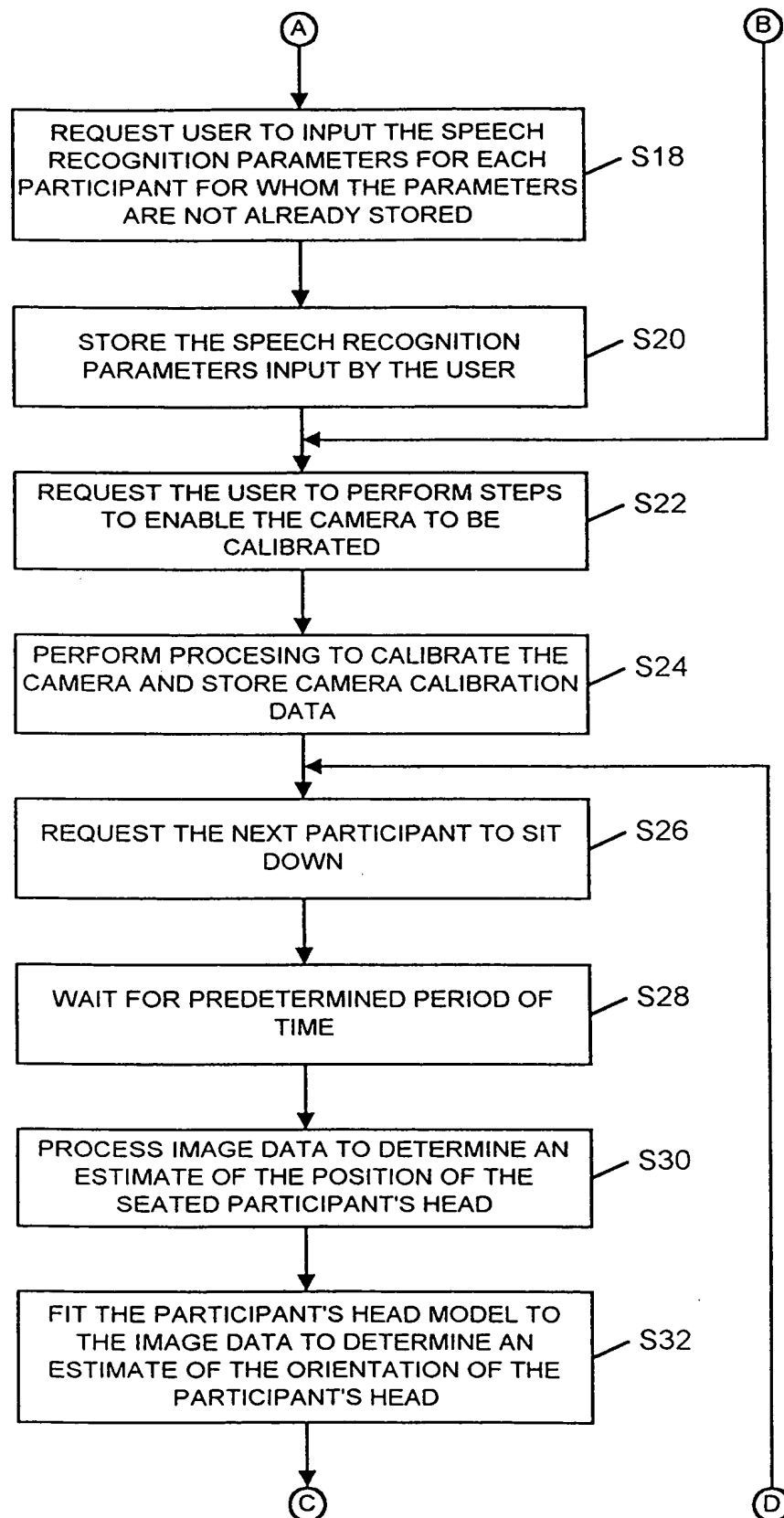


FIG. 3 (cont)

THIS PAGE BLANK (USPTO)

5/18

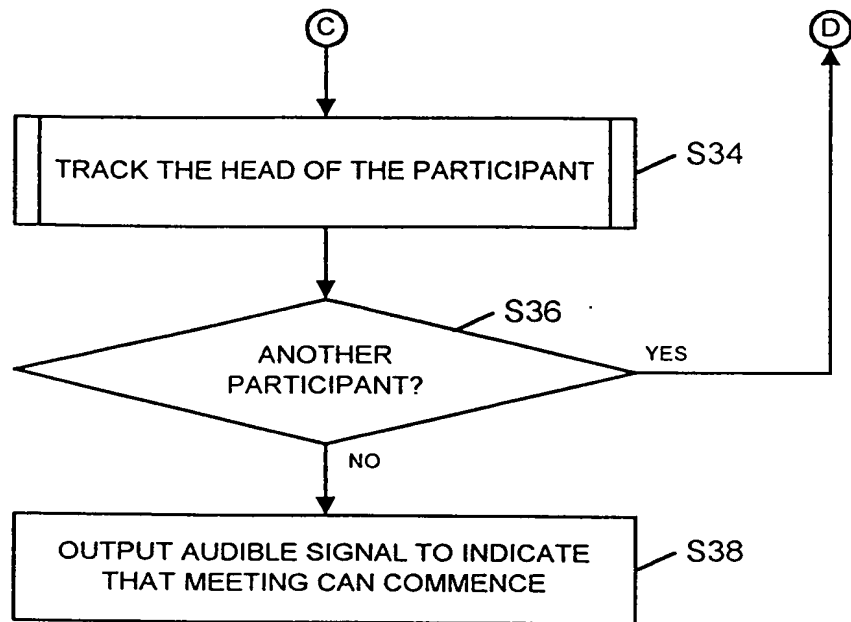


FIG. 3 (cont)

| NUMBER | PARTICIPANT |
|--------|-------------|
| 1 | MR. A |
| 2 | MR. B |
| 3 | MR. C |
| 4 | MISS. D |

FIG. 4

THIS PAGE BLANK (USPTO)

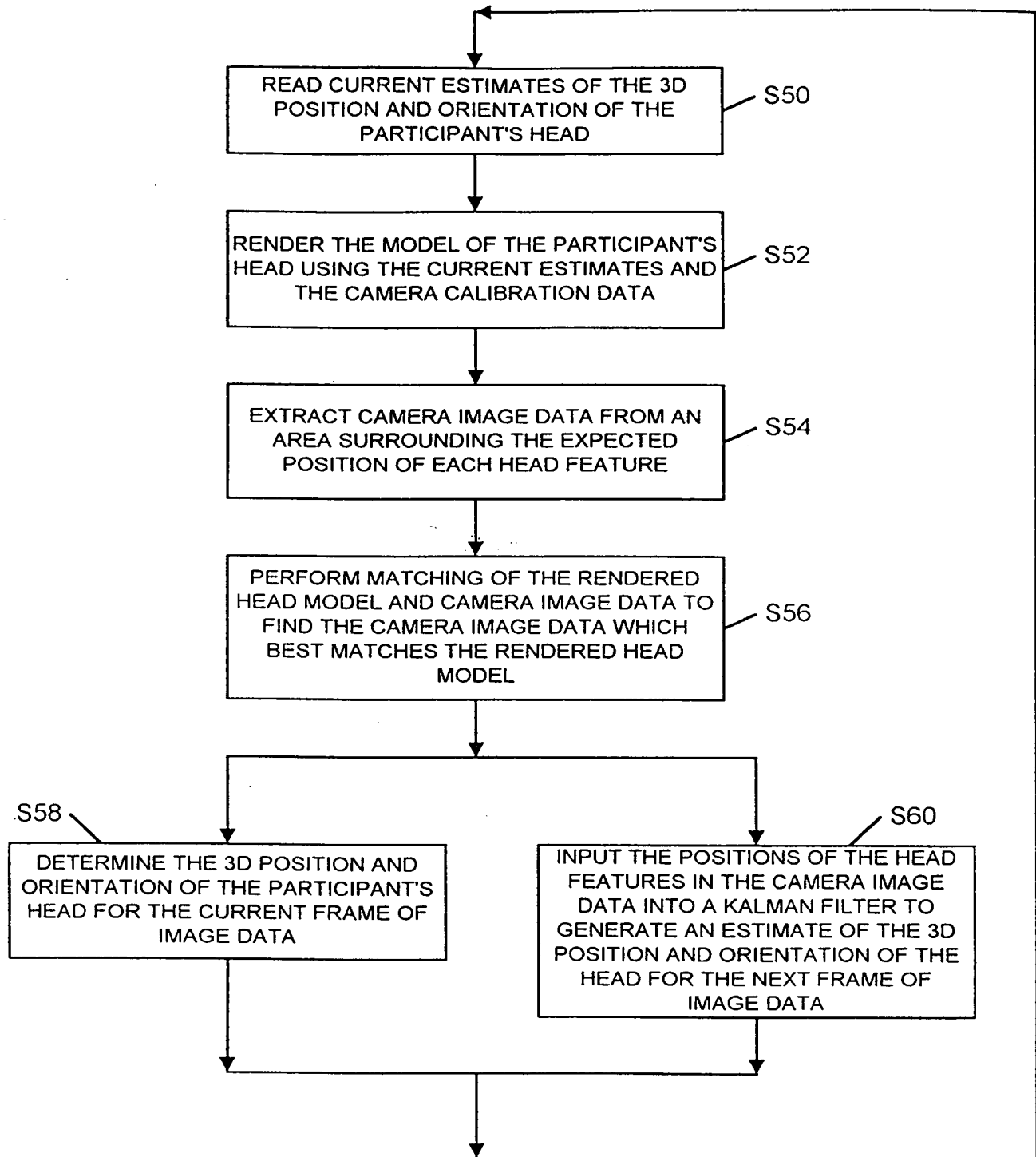
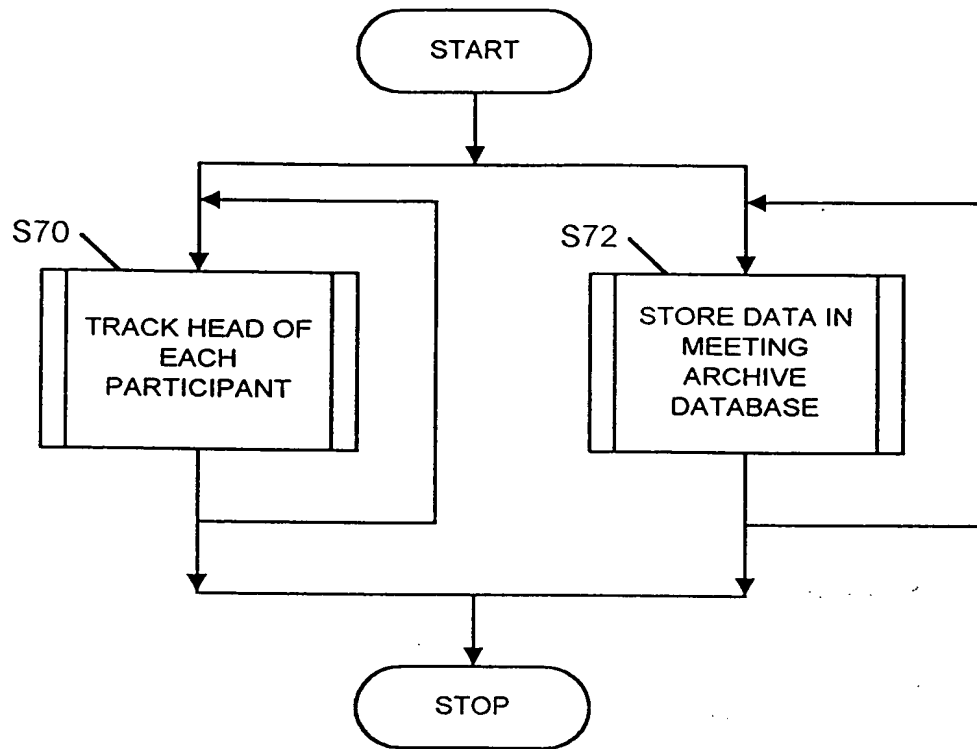


FIG. 5

THIS PAGE BLANK (USPTO)

*FIG. 6*

THIS PAGE BLANK (USPTO)

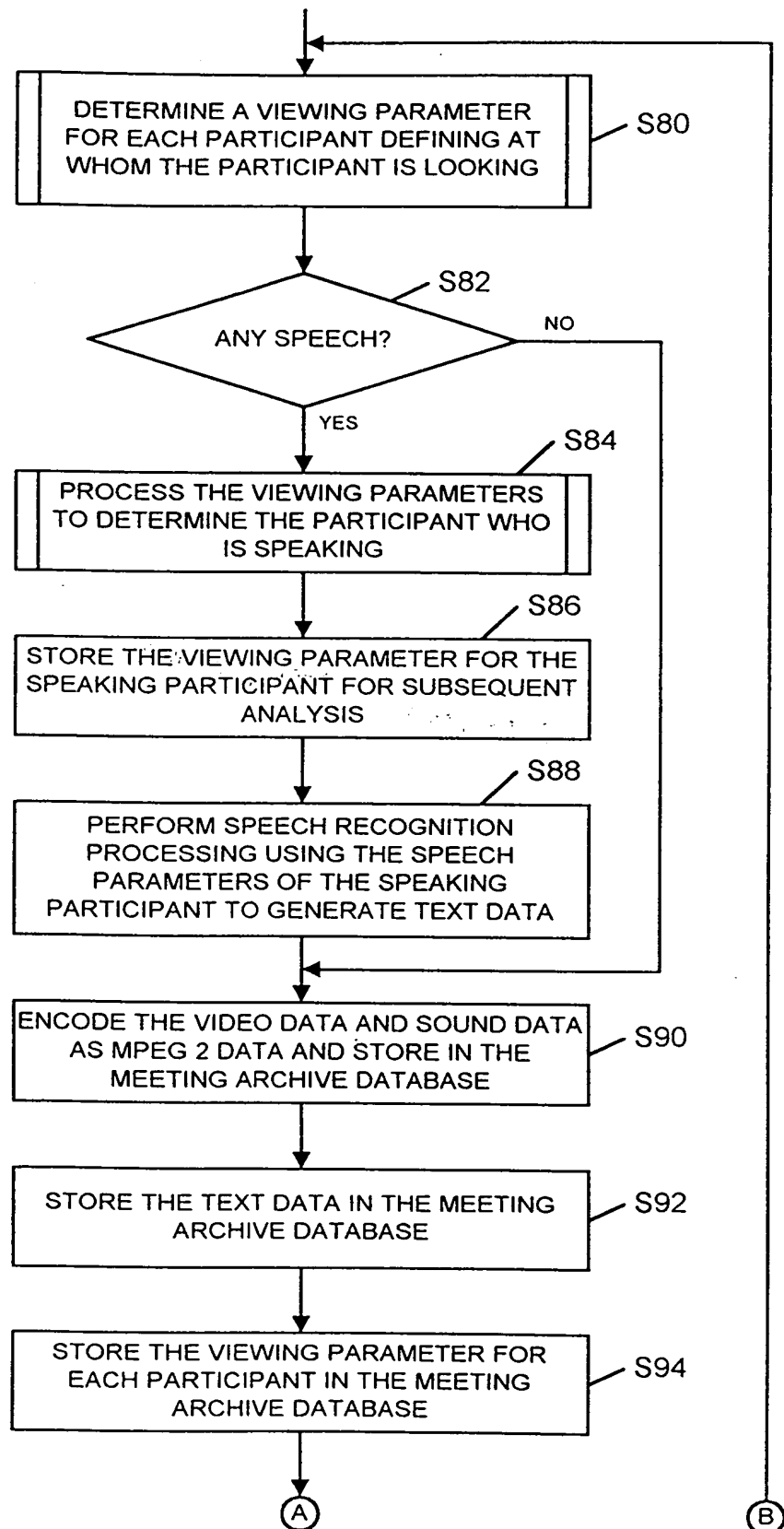


FIG. 7

THIS PAGE BLANK (USPTO)

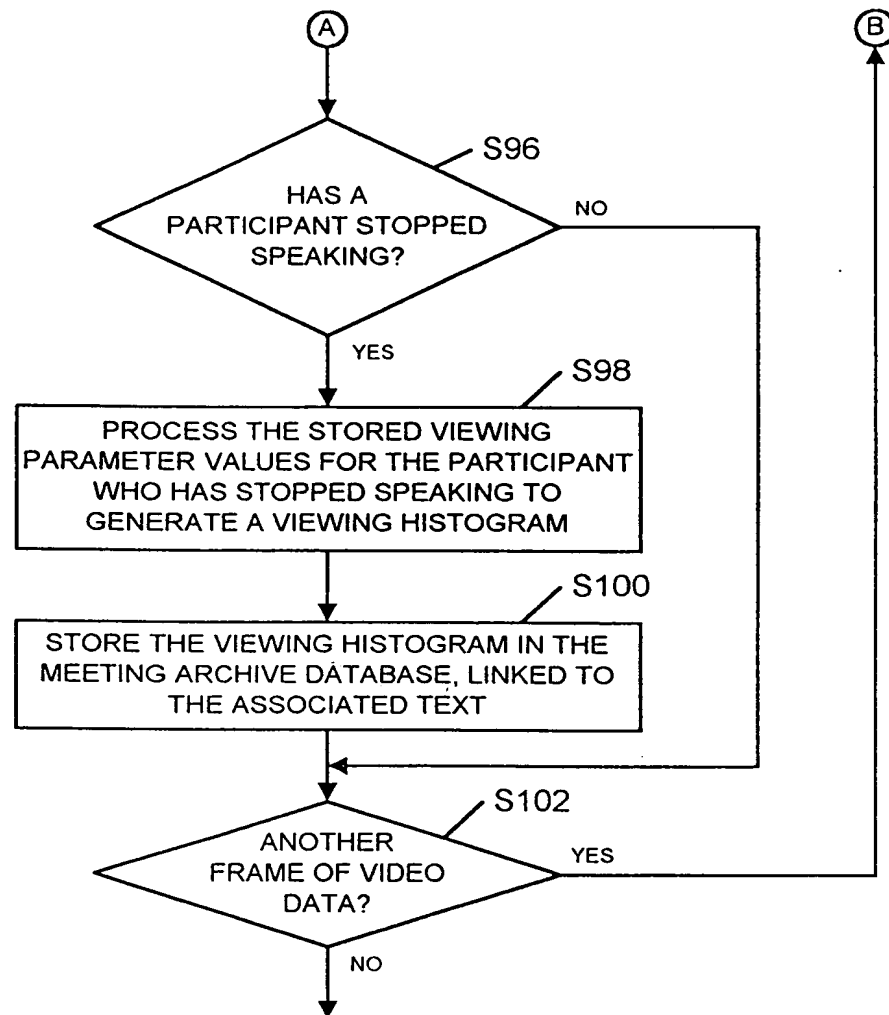


FIG. 7 (cont)

THIS PAGE BLANK (USPTO)

10/18

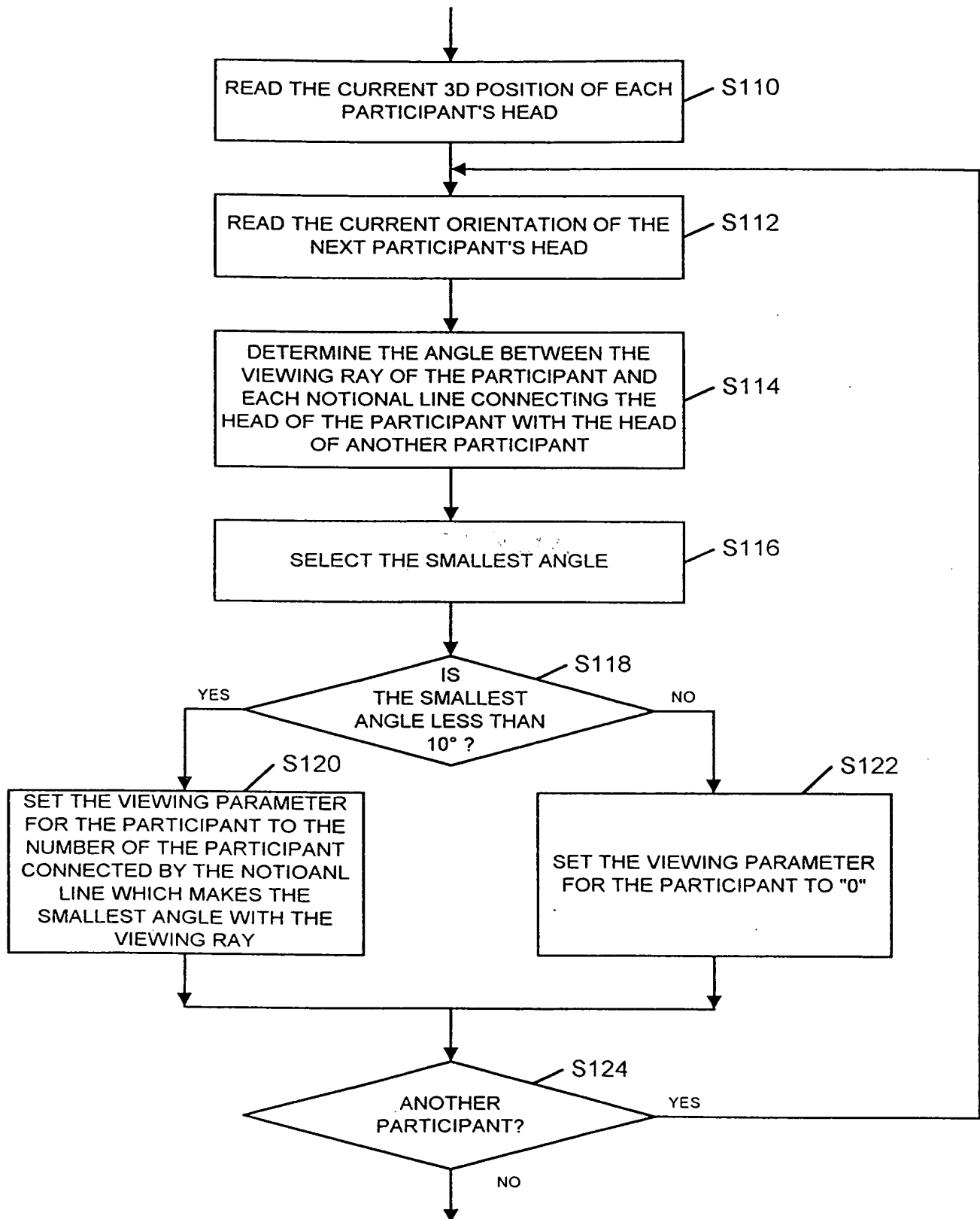
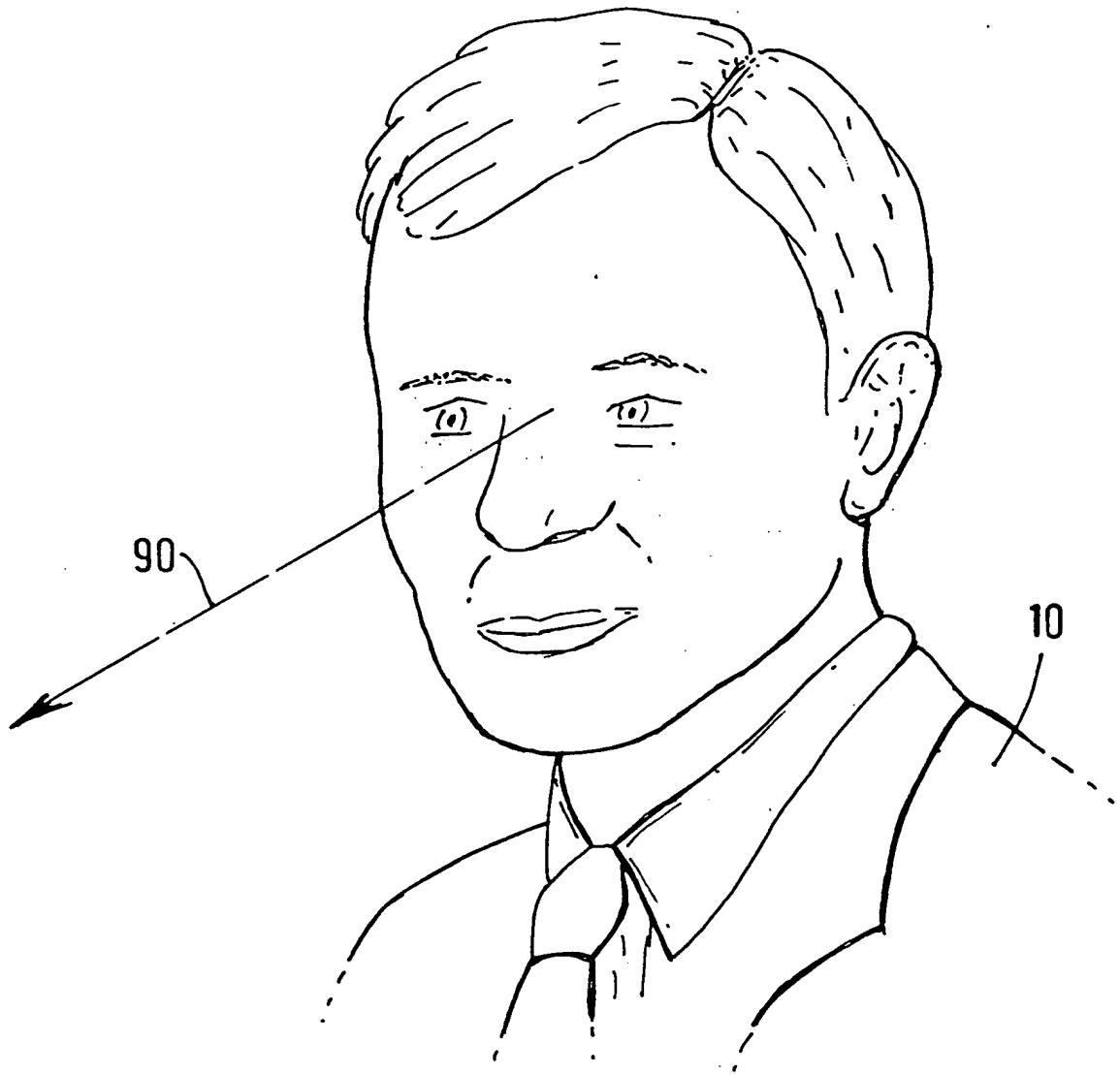


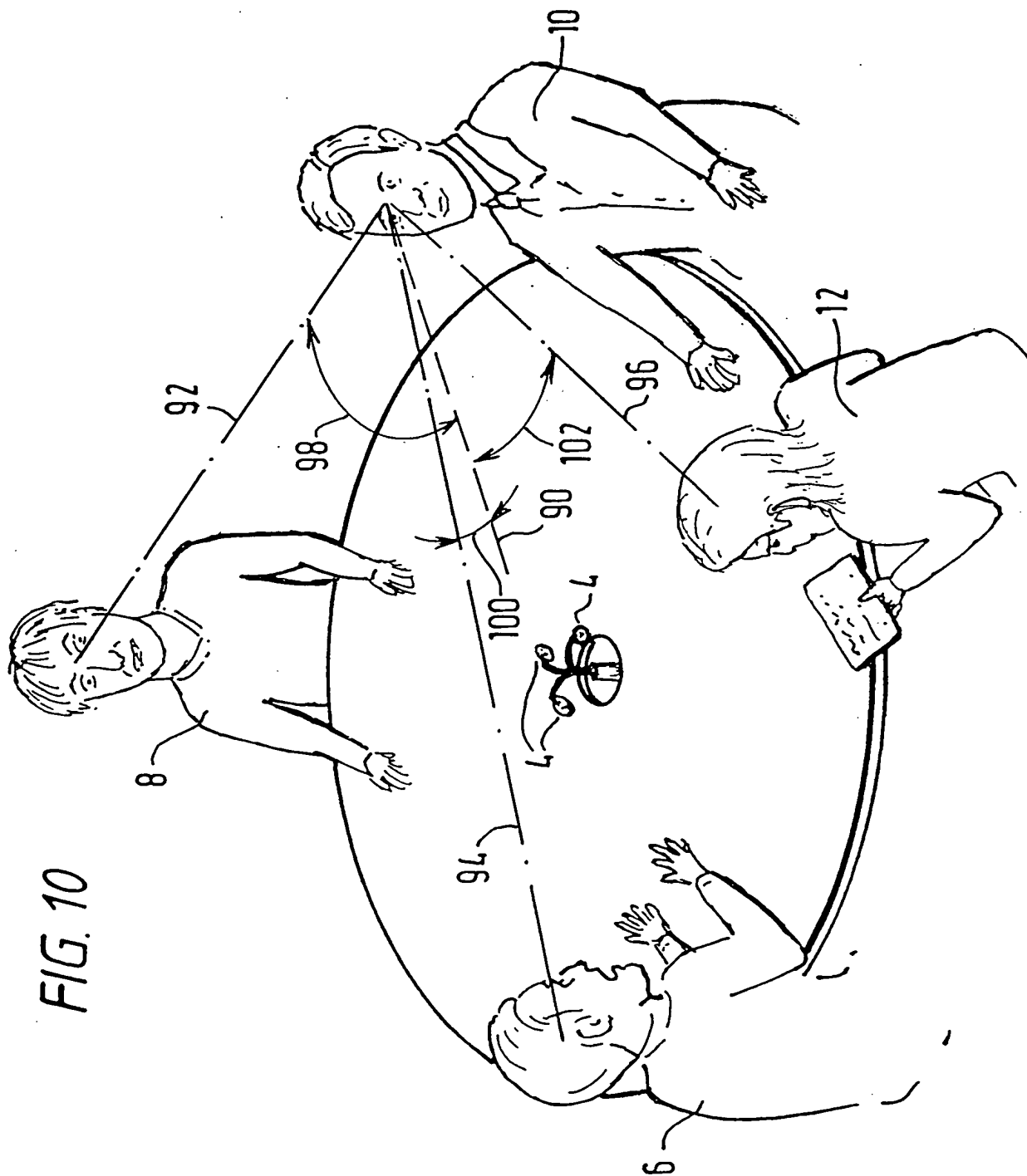
FIG. 8

THIS PAGE BLANK (USPTO)

FIG. 9



THIS PAGE BLANK (USPTO)



THIS PAGE BLANK (USPTO)

13/18

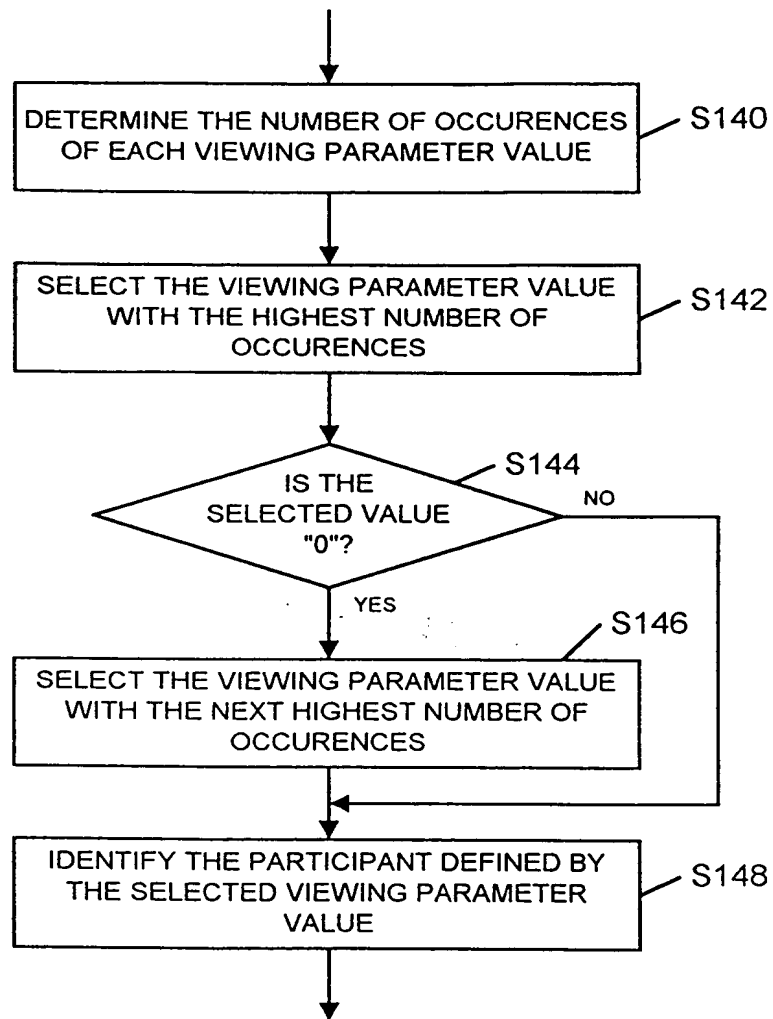
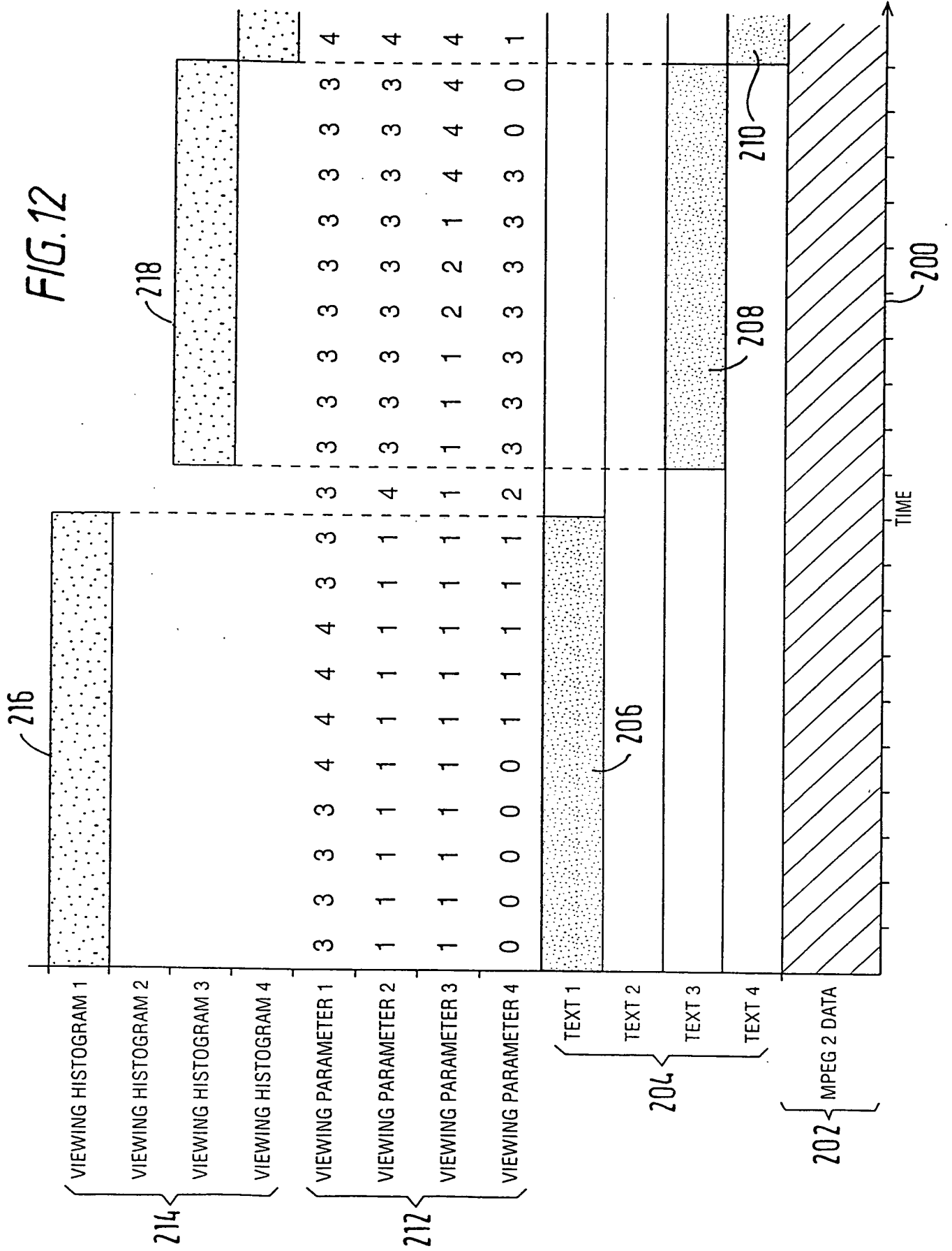


FIG. 11

THIS PAGE BLANK (USPTO)

FIG. 12



THIS PAGE BLANK (USPTO)

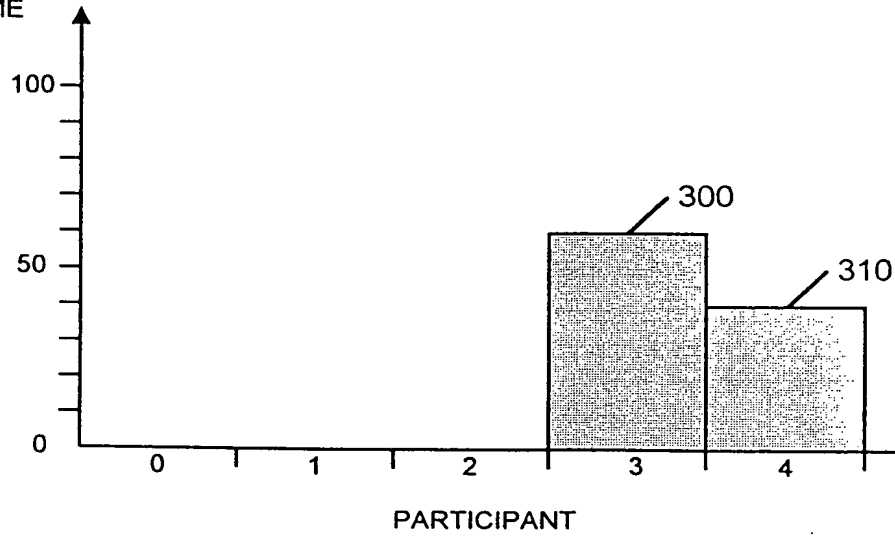
PERCENTAGE
GAZE TIME

FIG. 13A

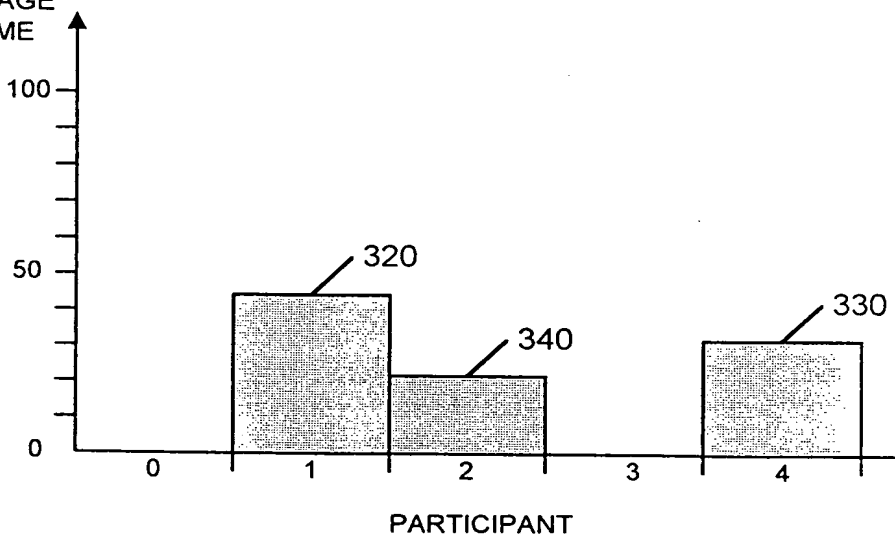
PERCENTAGE
GAZE TIME

FIG. 13B

THIS PAGE BLANK (USPTO)

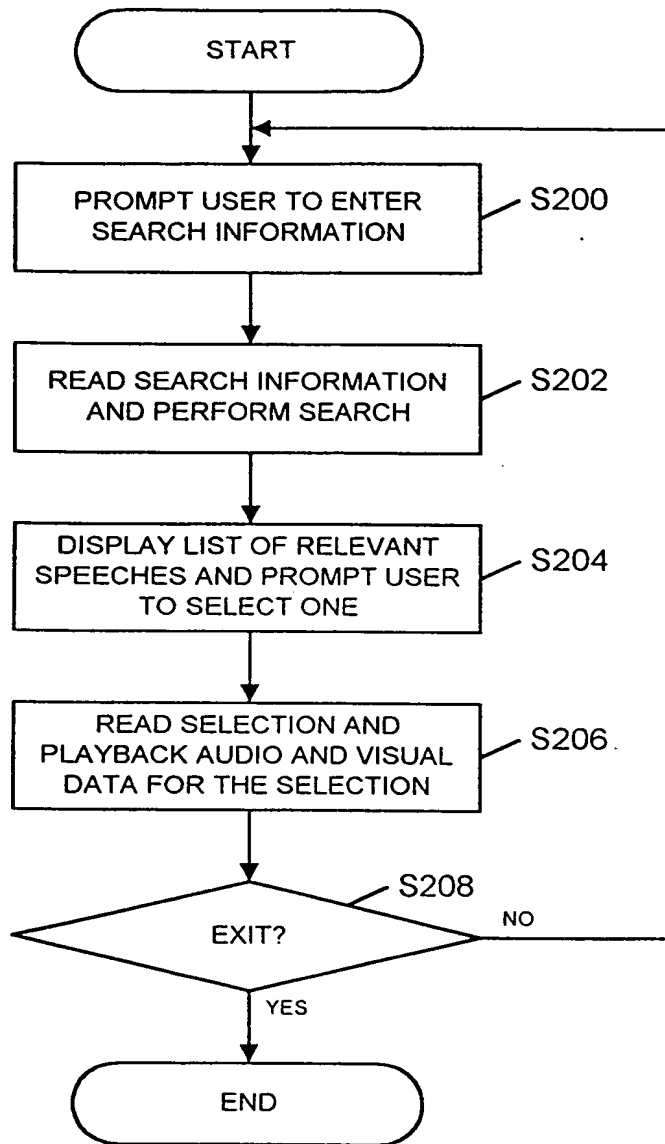


FIG. 14

THIS PAGE BLANK (USPTO)

Please enter search parameters

400 talking about 410 to 420

Time limits:

Before 430

After 440

Between 450 and 460

470

FIG. 15 A

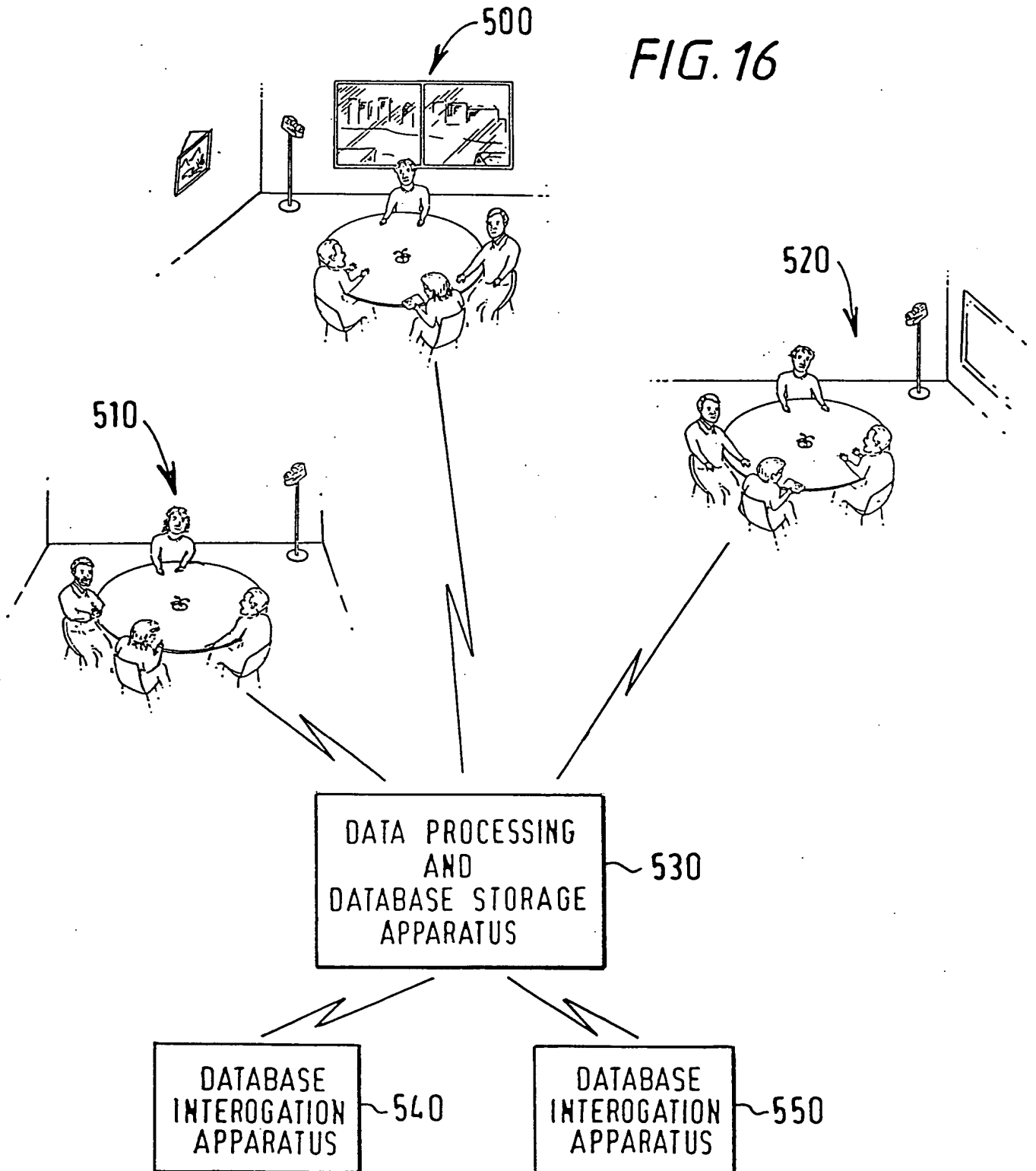
The following parts of the meeting are relevant. Please select one for playback:

1. Speech starting at 10 mins 0 secs (0.4 x full meeting time)
2. Speech starting at 12 mins 30 secs (0.5 x full meeting time)

FIG. 15 B

THIS PAGE BLANK (USPTO)

FIG. 16



THIS PAGE BLANK (USPTO)

**This Page is Inserted by IFW Indexing and Scanning
Operations and is not part of the Official Record**

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☐ BLACK BORDERS
- ☐ IMAGE CUT OFF AT TOP, BOTTOM OR SIDES
- ☒ FADED TEXT OR DRAWING
- ☒ BLURRED OR ILLEGIBLE TEXT OR DRAWING
- ☐ SKEWED/SLANTED IMAGES
- ☒ COLOR OR BLACK AND WHITE PHOTOGRAPHS
- ☐ GRAY SCALE DOCUMENTS
- ☒ LINES OR MARKS ON ORIGINAL DOCUMENT
- ☐ REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY
- ☐ OTHER: _____

IMAGES ARE BEST AVAILABLE COPY.

As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.

THIS PAGE BLANK (USPTO)